Finite Elements IV:
Exercises and solutions

Alexandre Ern Jean-Luc Guermond

May 13, 2021

Contents

Part I. Elements of functional analysis

1 Lebesgue spaces 1

2 Weak derivatives and Sobolev spaces 5

3 Traces and Poincaré inequalities 11

4 Distributions and duality in Sobolev spaces 17

Part II. Introduction to finite elements

5 Main ideas and definitions 21

6 One-dimensional finite elements and tensorization 27

7 Simplicial finite elements 35

Part III. Finite element interpolation

8 Meshes 43

9 Finite element generation 47

10 Mesh orientation 51

11 Local interpolation on affine meshes 55

12 Local inverse and functional inequalities 61

13 Local interpolation on nonaffine meshes 67

14 $\boldsymbol{H}(\mathrm{div})$ finite elements 71

15 $\boldsymbol{H}$(curl) finite elements **77**

16 Local interpolation in $\boldsymbol{H}$(div) and $\boldsymbol{H}$(curl) (I) **83**

17 Local interpolation in $\boldsymbol{H}$(div) and $\boldsymbol{H}$(curl) (II) **87**

Part IV. Finite element spaces

18 From broken to conforming spaces **93**

19 Main properties of the conforming subspaces **97**

20 Face gluing **101**

21 Construction of the connectivity classes **107**

22 Quasi-interpolation and best approximation **109**

23 Commuting quasi-interpolation **115**

Part V. Weak formulations and well-posedness

24 Weak formulation of model problems **121**

25 Main results on well-posedness **125**

Part VI. Galerkin approximation

26 Basic error analysis **133**

27 Error analysis with variational crimes **137**

28 Linear algebra **145**

29 Sparse matrices **153**

30 Quadratures **157**

Part VII. Elliptic PDEs: conforming approximation

31 Scalar second-order elliptic PDEs **163**

32 H^1-conforming approximation (I) **169**

33 H^1-conforming approximation (II) **175**

34 A posteriori error analysis **181**

35 The Helmholtz problem **185**

Part VIII. Elliptic PDEs: nonconforming approximation

36 Crouzeix–Raviart approximation **189**

37 Nitsche's boundary penalty method **193**

38 Discontinuous Galerkin **197**

39 Hybrid high-order method **203**

40 Contrasted diffusivity (I) **213**

41 Contrasted diffusivity (II) **217**

Part IX. Vector-valued elliptic PDEs

42 Linear elasticity **221**

43 Maxwell's equations: $\boldsymbol{H}(\mathrm{curl})$-approximation **227**

44 Maxwell's equations: control on the divergence **229**

45 Maxwell's equations: further topics **233**

Part X. Eigenvalue problems

46 Symmetric elliptic eigenvalue problems **237**

47 Symmetric operators, conforming approximation **247**

48 Nonsymmetric problems **251**

Part XI. PDEs in mixed form

49 Well-posedness for PDEs in mixed form **257**

50 Mixed finite element approximation **263**

51 Darcy's equations **271**

52 Potential and flux recovery **277**

53 Stokes equations: Basic ideas **283**

54 Stokes equations: Stable pairs (I) **287**

55 Stokes equations: Stable pairs (II) **291**

Part XII. First-order PDEs

56 Friedrichs' systems **299**

57 Residual-based stabilization **305**

58 Fluctuation-based stabilization (I) **309**

59 Fluctuation-based stabilization (II) **313**

60 Discontinuous Galerkin **317**

61 Advection-diffusion **321**

62 Stokes equations: Residual-based stabilization **327**

63 Stokes equations: Other stabilizations **333**

Part XIII. Parabolic PDEs

64 Bochner integration **339**

65 Weak formulation and well-posedness **345**

66 Semi-discretization in space **349**

67 Implicit and explicit Euler schemes **353**

68 BDF2 and Crank–Nicolson schemes **357**

69 Discontinuous Galerkin in time **363**

70 Continuous Petrov–Galerkin in time **371**

71 Analysis using inf-sup stability **377**

Part XIV. Time-dependent Stokes equations

72 Weak formulations and well-posedness **387**

73 Monolithic time discretization **391**

74 Projection methods **393**

75 Artificial compressibility **399**

Part XV. Time-dependent linear PDEs

76 Well-posedness and space semi-discretization **407**

77 Implicit time discretization **411**

78 Explicit time discretization **417**

Part XVI. Nonlinear hyperbolic PDEs

79 Scalar conservation equations **427**

80 Hyperbolic systems **433**

81 First-order approximation **441**

82 Higher-order approximation **445**

83 Higher-order approximation and limiting **453**

Chapter 1

Lebesgue spaces

Exercises

Exercise 1.1 (Measurability). Let W be a nonmeasurable subset of $D := (0,1)$. Let $f : W \to \mathbb{R}$ be defined by $f(x) := 1$ if $x \in D \backslash W$ and $f(x) := 0$ if $x \in W$. (i) Is f measurable? (ii) Assume that there is a measurable subset $V \subset W$ s.t. $|V| > 0$. Compute $\sup_{x\in D} f(x)$, $\operatorname{ess\,sup}_{x\in D} f(x)$, $\inf_{x\in D} f(x)$, $\operatorname{ess\,inf}_{x\in D} f(x)$. (iii) Is f a member of $L^\infty(D)$? (iv) Assume now that W has zero measure (hence, W is measurable). Compute $\inf_{x\in D} f(x)$ and $\operatorname{ess\,inf}_{x\in D} f(x)$.

Exercise 1.2 (Measurability and equality a.e.). Prove Corollary 1.11. (*Hint*: consider the sets $A_r := \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) > r\}$ and $B_r := \{\boldsymbol{x} \in D \mid g(\boldsymbol{x}) > r\}$ for all $r \in \mathbb{R}$, and show that $B_r = (A_r \cap (A_r \backslash B_r)^c) \cup (B_r \backslash A_r)$.)

Exercise 1.3 (Lebesgue's theorem). Let $D := (-1,1)$. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions in $L^1(D)$ and let $g \in L^1(D)$. Assume that $f_n \to f$ a.e. in D. Propose a counterexample to show that the assumption "$|f_n| \leq g$ a.e. for all $n \in \mathbb{N}$" cannot be replaced by "$f_n \leq g$ a.e. for all $n \in \mathbb{N}$" in Lebesgue's dominated convergence theorem.

Exercise 1.4 (Compact support). Let $D := (0,1)$ and $f(x) := 1$ for all $x \in D$. What is the support of f in D? Is the support compact?

Exercise 1.5 (Pointwise limit of measurable functions). Let D be a measurable set in $\mathbb{R}^d$. Let $f_n : D \to \mathbb{R}$ for all $n \in \mathbb{N}$ be real-valued measurable functions. (i) Show that $\limsup_{n\in\mathbb{N}} f_n$ and $\liminf_{n\in\mathbb{N}} f_n$ are both measurable. (*Hint*: recall that $\limsup_{n\in\mathbb{N}} f_n(\boldsymbol{x}) := \inf_{n\in\mathbb{N}} \sup_{k\geq n} f_k(\boldsymbol{x})$ and $\liminf_{n\in\mathbb{N}} f_n(\boldsymbol{x}) := \sup_{n\in\mathbb{N}} \inf_{k\geq n} f_k(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$). (ii) Let $f : D \to \mathbb{R}$. Assume that $f_n(\boldsymbol{x}) \to f(\boldsymbol{x})$ for every $\boldsymbol{x} \in D$. Show that f is measurable. (iii) Let $f : D \to \mathbb{R}$. Assume that $f_n(\boldsymbol{x}) \to f(\boldsymbol{x})$ for a.e. $\boldsymbol{x} \in D$. Show that f is measurable.

Exercise 1.6 (Operations on measurable functions). The objective of this exercise is to prove Theorem 1.6. Let $f : D \to \mathbb{R}$ and $g : D \to \mathbb{R}$ be two measurable functions and let $\lambda \in \mathbb{R}$. (i) Show that λf is measurable. (*Hint*: use Lemma 1.9). (ii) Idem for $|f|$. (iii) Idem for $f + g$. (iv) Idem for fg. (*Hint*: observe that $fg = \frac{1}{2}(f+g)^2 - \frac{1}{2}(f-g)^2$.)

Solution to exercises

Exercise 1.1 (Measurability). (i) Since the set $\{x \in D \mid f(x) < 1\} = W$ is not measurable, Lemma 1.9 implies that f is not measurable.
(ii) We have

$$\begin{aligned}\sup_{x\in D} f(x) &= 1,\\ \operatorname*{ess\,sup}_{x\in D} f(x) &= 1,\\ \inf_{x\in D} f(x) &= 0,\\ \operatorname*{ess\,inf}_{x\in D} f(x) &= 0.\end{aligned}$$

(iii) Although $\operatorname{ess\,sup}_{x\in D} |f(x)| = 1 \leq \infty$, the function f is not a member of $L^\infty(D)$ since it is not measurable.
(iv) Since we now assume that $|W| = 0$, i.e., $f = 1$ a.e. in D, we have

$$\begin{aligned}\inf_{x\in D} f(x) &= 0,\\ \operatorname*{ess\,inf}_{x\in D} f(x) &= 1.\end{aligned}$$

Exercise 1.2 (Measurability and equality a.e.). Following the hint, let $A_r := \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) > r\}$ and $B_r := \{\boldsymbol{x} \in D \mid g(\boldsymbol{x}) > r\}$ for all $r \in \mathbb{R}$. By assumption, the set A_r is measurable. We observe that $A_r \backslash B_r \subset \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) \neq g(\boldsymbol{x})\}$ and $B_r \backslash A_r \subset \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) \neq g(\boldsymbol{x})\}$. Hence, $|A_r \backslash B_r|^* = 0$ and $|B_r \backslash A_r|^* = 0$. This means that $A_r \backslash B_r$ and $B_r \backslash A_r$ are measurable (see Example 1.4). After observing that $B_r \cap A_r = A_r \cap (A_r \backslash B_r)^c$ and $B_r \cap A_r^c = B_r \backslash A_r$, we finally have $B_r = (A_r \cap (A_r \backslash B_r)^c) \cup (B_r \backslash A_r)$. This shows that B_r is measurable since the sets A_r, $(A_r \backslash B_r)^c$, and $B_r \backslash A_r$ are measurable. Hence, g is measurable owing to Lemma 1.9.

Exercise 1.3 (Lebesgue's theorem). The sequence $\{f_n\}_{n\geq 1}$ such that $f_n(x) := -2n$ if $|x| \leq \frac{1}{n}$ and $f_n(x) := 0$ otherwise is such that $f_n \to 0$ a.e. in D and $f_n \leq 0 \in L^1(D)$, but f_n does not converge to 0 in $L^1(D)$ since $\|f_n\|_{L^1(D)} = 1$.

Exercise 1.4 (Compact support). We have $\{x \in D \mid f(x) \neq 0\} = D$. The slight subtlety here is that the closure of D in D is D itself. Hence, the support of f in D is D. Note that D is not compact since it is not a closed set in $\mathbb{R}$ (the limit point of the sequence $\{\frac{1}{n}\}_{n\geq 1}$ does not belong to D).

Exercise 1.5 (Pointwise limit of measurable functions). (i) Using the hint, we have for all $\boldsymbol{x} \in D$,

$$\begin{aligned}\limsup_{n\in\mathbb{N}} f_n(\boldsymbol{x}) \leq c &\iff \inf_{n\in\mathbb{N}} \sup_{k\geq n} f_k(\boldsymbol{x}) \leq c\\ &\iff \forall j \geq 1,\ \exists n \geq 0, \quad \sup_{k\geq n} f_k(\boldsymbol{x}) \leq c + \frac{1}{j}\\ &\iff \forall j \geq 1,\ \exists n \geq 0,\ \forall k \geq n, \quad f_k(\boldsymbol{x}) \leq c + \frac{1}{j}\\ &\iff \boldsymbol{x} \in \bigcap_{j\geq 1} \bigcup_{n\geq 0} \bigcap_{k\geq n} \Big\{\boldsymbol{y} \in D \mid f(\boldsymbol{y}) \leq c + \frac{1}{j}\Big\}.\end{aligned}$$

This proves that

$$\Big\{\boldsymbol{x} \in D \mid \limsup_{n\in\mathbb{N}} f_n(\boldsymbol{x}) \leq c\Big\} = \bigcap_{j\geq 1} \bigcup_{n\geq 0} \bigcap_{k\geq n} \Big\{\boldsymbol{y} \in D \mid f(\boldsymbol{y}) \leq c + \frac{1}{j}\Big\}.$$

Hence, the function $\limsup_{n\in\mathbb{N}} f_n$ is measurable. The proof that $\liminf_{n\in\mathbb{N}} f_n$ is a measurable function is similar.
(ii) Saying that $f_n(\boldsymbol{x}) \to f(\boldsymbol{x})$ for every $\boldsymbol{x} \in D$ means that

$$\limsup_{n\in\mathbb{N}} f_n(\boldsymbol{x}) = f(\boldsymbol{x}) = \liminf_{n\in\mathbb{N}} f_n(\boldsymbol{x})$$

for every $\boldsymbol{x} \in D$. We conclude from Step (i) that f is measurable.
(iii) Let $S \subset D$ be such that $S := \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) = \limsup_{n\in\mathbb{N}} f_n(\boldsymbol{x}) = \liminf_{n\in\mathbb{N}} f_n(\boldsymbol{x})\}$. By assumption, we have $|S^c| = 0$. Moreover, we have

$$\begin{aligned} S &= \{\boldsymbol{x} \in D \mid \limsup_{n\in\mathbb{N}} f_n(\boldsymbol{x}) = \liminf_{n\in\mathbb{N}} f_n(\boldsymbol{x}) = f(\boldsymbol{x})\} \\ &\subset \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) = \limsup f_n(\boldsymbol{x})\}. \end{aligned}$$

Hence, $\{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) \neq \limsup f_n(\boldsymbol{x})\} \subset S^c$. This means that the function f and $\limsup f_n$ coincide almost everywhere. Corollary 1.11 implies that f is measurable since $\limsup f_n$ is measurable.

Exercise 1.6 (Operations on measurable functions). (i) There is nothing to prove if $\lambda = 0$. Assume now that $\lambda > 0$. For all $r \in \mathbb{R}$, we have

$$\{\boldsymbol{x} \in D \mid \lambda f(\boldsymbol{x}) > r\} = \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) > r/\lambda\}.$$

Hence, $\{\boldsymbol{x} \in D \mid \lambda f(\boldsymbol{x}) > r\}$ is measurable. The reasoning for $\lambda < 0$ is similar. We conclude that λf is measurable by invoking Lemma 1.9.
(ii) For all $r \in \mathbb{R}$, we have

$$\{\boldsymbol{x} \in D \mid |f(\boldsymbol{x})| > r\} = \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) > r\} \bigcup \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) < -r\}.$$

Hence, $\{\boldsymbol{x} \in D \mid |f(\boldsymbol{x})| > r\}$ is measurable (recall that the union of two measurable sets is measurable). We conclude by using Lemma 1.9.
(ii) Recall that if $f(\boldsymbol{x}) > r - g(\boldsymbol{x})$, there exists $q \in \mathbb{Q}$ such that $f(\boldsymbol{x}) > q > r - g(\boldsymbol{x})$. Then, for all $r \in \mathbb{R}$, we have

$$\begin{aligned} \{\boldsymbol{x} \in D \mid (f+g)(\boldsymbol{x}) > r\} &= \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) > r - g(\boldsymbol{x})\} \\ &= \bigcup_{q\in\mathbb{Q}} \{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) > q\} \cap \{\boldsymbol{x} \in D \mid g(\boldsymbol{x}) > r - q\}. \end{aligned}$$

Since any countable union of measurable sets is measurable, Lemma 1.9 implies that $f + g$ is measurable.
(ii) Let us notice that $fg = \frac{1}{2}(f+g)^2 - \frac{1}{2}(f-g)^2$. Hence, Theorem 1.14 combined with (ii) implies that fg is measurable.

Chapter 2

Weak derivatives and Sobolev spaces

Exercises

Exercise 2.1 (Lebesgue point). Let $a \in \mathbb{R}$. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) := 0$ if $x < 0$, $f(0) := a$, and $f(x) := 1$ if $x > 0$. Show that 0 is not a Lebesgue point of f for all a.

Exercise 2.2 (Lebesgue differentiation). The goal is to prove Theorem 2.2. (i) Let $h \in \mathcal{H}$ (the sign of h is unspecified). Show that $R(x,h) := \frac{F(x+h)-F(x)}{h} - f(x) = \frac{1}{h}\int_x^{x+h}(f(t)-f(x))\,\mathrm{d}t$. (ii) Conclude.

Exercise 2.3 (Lebesgue measure and weak derivative). Let $D := (0,1)$. Let C_∞ be the Cantor set (see Example 1.5). Let $f : D \to \mathbb{R}$ be defined by $f(x) := x$ if $x \notin C_\infty$, and $f(x) := 2 - 5x$ if $x \in C_\infty$. (i) Is f measurable? (*Hint*: use Corollary 1.11.) (ii) Compute $\sup_{x\in D} f(x)$, $\operatorname{ess\,sup}_{x\in D} f(x)$, $\inf_{x\in D} f(x)$, $\operatorname{ess\,inf}_{x\in D} f(x)$, and $\|f\|_{L^\infty(D)}$. (iii) Show that f is weakly differentiable and compute $\partial_x f(x)$. (iv) Compute $f(x) - \int_0^x \partial_t f(t)\,\mathrm{d}t$ for all $x \in D$. (iv) Identify the function $f^{\mathrm{c}} \in C^0(\overline{D})$ that satisfies $f = f^{\mathrm{c}}$ a.e. on D? Compute $f^{\mathrm{c}}(x) - \int_0^x \partial_t f(t)\,\mathrm{d}t$ for all $x \in D$.

Exercise 2.4 (Weak derivative). Let $D := (-1,1)$. Prove that if $u \in L^1_{\mathrm{loc}}(D)$ has a second-order weak derivative, it also has a first-order weak derivative. (*Hint*: consider $\psi(x) := \int_{-1}^x (\varphi(t) - c_\varphi \rho(t))\,\mathrm{d}t$ for all $\varphi \in C_0^\infty(D)$, with $c_\varphi := \int_D \varphi\,\mathrm{d}x$, $\rho \in C_0^\infty(D)$, and $\int_D \rho\,\mathrm{d}x = 1$.)

Exercise 2.5 (Clairaut's theorem). Let $v \in L^1_{\mathrm{loc}}(D)$. Let $\alpha, \beta \in \mathbb{N}^d$ and assume that the weak derivatives $\partial^\alpha v$, $\partial^\beta v$ exist and that the weak derivative $\partial^\alpha(\partial^\beta v)$ exists. Prove that $\partial^\beta(\partial^\alpha v)$ exists and $\partial^\alpha(\partial^\beta v) = \partial^\beta(\partial^\alpha v)$.

Exercise 2.6 (Weak and classical derivatives). Let $k \in \mathbb{N}$, $k \geq 1$, and let $v \in C^k(D)$. Prove that, up to the order k, the weak derivatives and the classical derivatives of v coincide.

Exercise 2.7 ($H^1(D)$). (i) Let $D := (-1,1)$ and $u : D \to \mathbb{R}$ s.t. $u(x) := |x|^{\frac{3}{2}} - 1$. Determine whether u is a member of $H^1(D;\mathbb{R})$. (ii) Let $u_1 \in C^1((-1,0];\mathbb{R})$ and $u_2 \in C^1([0,1);\mathbb{R})$ and assume that $u_1(0) = u_2(0)$. Let u be such that $u_{|(-1,0)} := u_1$ and $u_{|(0,1)} := u_2$. Determine whether u is a member of $H^1(D;\mathbb{R})$. Explain why $u \notin H^1(D;\mathbb{R})$ if $u_1(0) \neq u_2(0)$.

Exercise 2.8 (Broken seminorm). Let D be an open set in $\mathbb{R}^d$. Let $\{D_1, \dots, D_n\}$ be a partition of D as in Remark 2.13. (i) Show that $(\nabla v)_{|D_i} = \nabla(v_{|D_i})$ for all $i \in \{1\!:\!n\}$ and all $v \in W^{1,1}_{\mathrm{loc}}(D)$. (ii) Let $p \in [1,\infty)$ and $v \in W^{1,p}(D)$. Show that $\sum_{i\in\{1:n\}} |v_{|D_i}|^p_{W^{1,p}(D_i)} = |v|^p_{W^{1,p}(D)}$. (iii) Let $s \in (0,1)$, $p \in [1,\infty)$, and $v \in W^{s,p}(D)$. Prove that $\sum_{i\in\{1:n\}} |v_{|D_i}|^p_{W^{s,p}(D_i)} \le |v|^p_{W^{s,p}(D)}$.

Exercise 2.9 ($W^{s,p}$). Let D be a bounded open set in $\mathbb{R}^d$. Let $\alpha \in (0,1]$. Show that $C^{0,\alpha}(D;\mathbb{R}) \hookrightarrow W^{s,p}(D;\mathbb{R})$ for all $p \in [1,\infty)$ if $s \in [0,\alpha)$.

Exercise 2.10 (Unbounded function in $H^1(D)$). Let $D := B(\mathbf{0}, \frac{1}{2}) \subset \mathbb{R}^2$ be the ball centered at 0 and of radius $\frac{1}{2}$. (i) Show that the (unbounded) function $u(\boldsymbol{x}) := \ln\big(-\ln(\|\boldsymbol{x}\|_{\ell^2})\big)$ has weak partial derivatives. (*Hint*: work on $D\backslash B(\mathbf{0},\epsilon)$ with $\epsilon \in (0,\frac{1}{2})$, and use Lebesgue's dominated convergence theorem.) (ii) Show that u is in $H^1(D)$.

Exercise 2.11 (Equivalent norm). Let $m \in \mathbb{N}$, $m \ge 2$, and let $p \in [1,\infty)$. Prove that the norm $\|v\| := (\|v\|^p_{L^p} + \ell_D^{mp}|v|^p_{W^{m,p}(D)})^{\frac{1}{p}}$ is equivalent to the canonical norm in $W^{m,p}(D)$. (*Hint*: use the Peetre–Tartar lemma (Lemma A.20) and invoke the compact embeddings from Theorem 2.35.)

Solution to exercises

Exercise 2.1 (Lebesgue point). Let $r > 0$. We have

$$\frac{1}{2r}\int_{-r}^{r} |f(t) - a| \,\mathrm{d}t = \frac{1}{2r}\int_{-r}^{0} |a| \,\mathrm{d}t + \frac{1}{2r}\int_{0}^{r} |1 - a| \,\mathrm{d}t = \frac{1}{2}(|a| + |1-a|).$$

Since $\frac{1}{2}(|a| + |1-a|) \ge \frac{1}{2}$ for all $a \in \mathbb{R}$, this proves that $\frac{1}{2r}\int_{-r}^{r} |f(t) - a| \,\mathrm{d}t$ cannot converge to zero a $r \downarrow 0$. Hence, 0 is not a Lebesgue point of f.

Exercise 2.2 (Lebesgue differentiation). (i) Let $x \in \mathbb{R}$. We have

$$\frac{F(x+h) - F(x)}{h} = \frac{1}{h}\int_{x}^{x+h} f(t) \,\mathrm{d}t,$$

for all $h \in \mathcal{H}$. We infer that

$$R(x,h) = \frac{1}{h}\int_{x}^{x+h} f(t) \,\mathrm{d}t - f(x) = \frac{1}{h}\int_{x}^{x+h} (f(t) - f(x)) \,\mathrm{d}t.$$

(ii) We want to prove that we have $|\frac{F(x+h)-F(x)}{h} - f(x)| \to 0$ as $h \to 0$, for every Lebesgue point x of f, i.e., $R(x,h) \to 0$. Recalling that the sign of h is unspecified and using the above expression for $R(x,h)$, we have

$$|R(x,h)| \le \frac{1}{|h|}\int_{x}^{x+h} |f(t) - f(x)| \,\mathrm{d}t \le 2\frac{1}{|2h|}\int_{x-h}^{x+h} |f(t) - f(x)| \,\mathrm{d}t,$$

which shows that $\lim_{h\to 0} |R(x,h)| = 0$ since x is a Lebesgue point of f. Hence, F is strongly differentiable at x.

Exercise 2.3 (Lebesgue measure and weak derivative). (i) Yes, f is measurable according to Corollary 1.11, since $f(x) = x$ for a.e. $x \in D$.
(ii) We have

$$\begin{aligned}
\sup_{x\in D} f(x) &= \max\left(\sup_{x\in D\setminus C_\infty} x, \sup_{x\in C_\infty} (2-5x)\right) = 2,\\
\operatorname*{ess\,sup}_{x\in D} f(x) &= \sup_{x\in D\setminus C_\infty} x = 1,\\
\inf_{x\in D} f(x) &= \min\left(\inf_{x\in D\setminus C_\infty} x, \inf_{x\in C_\infty} (2-5x)\right) = -3,\\
\operatorname*{ess\,inf}_{x\in D} f(x) &= \inf_{x\in D\setminus C_\infty} x = 0,\\
\|f\|_{L^\infty(D)} &= \operatorname*{ess\,sup}_{x\in D} |f(x)| = \sup_{x\in D\setminus C_\infty} |x| = 1.
\end{aligned}$$

(iii) For all $\phi \in C^\infty(D)$, we have

$$\int_D f(x)\partial_x\phi(x)\,\mathrm{d}x = \int_D x\partial_x\phi(x)\,\mathrm{d}x = \int_D -\phi(x)\,\mathrm{d}x.$$

Hence, f is weakly differentiable and $\partial_x f(x) = 1$ for a.e. $x \in D$.
(iv) For all $x \in D\setminus C_\infty$, we have

$$f(x) - \int_0^x \partial_t f(t)\,\mathrm{d}t = x - x = 0.$$

For all $x \in C_\infty$, we have

$$f(x) - \int_0^x \partial_t f(t)\,\mathrm{d}t = 2 - 5x - x = 2 - 6x.$$

Hence, the fundamental theorem of calculus for f holds true only a.e. on D.
(v) We have $f(x) = x$ for a.e. $x \in D$, hence $f^{\mathrm{c}}(x) = x$. (Observe that, in accordance with Theorem 2.26 with $d = 1$ and all $p = 1$, we indeed have $f^{\mathrm{c}} \in C^0(\overline{D})$.) Since $\partial_t f^{\mathrm{c}} = \partial_t f$ a.e. in D, the fundamental theorem of calculus implies that

$$f^{\mathrm{c}}(x) - \int_0^x \partial_t f(t)\,\mathrm{d}t = f^{\mathrm{c}}(x) - \int_0^x \partial_t f^{\mathrm{c}}(t)\,\mathrm{d}t = f^{\mathrm{c}}(0) = 0.$$

Exercise 2.4 (Weak derivative). Let $\rho \in C_0^\infty(D)$ with $\int_D \rho\,\mathrm{d}x = 1$. For all $\varphi \in C_0^\infty(D)$, the function ψ in the hint is in $C_0^\infty(D)$ and $\partial_x\psi = \varphi - c_\varphi\rho$ with $c_\varphi := \int_D \varphi\,\mathrm{d}x$. Letting $v := \partial_{xx}u$ and $C_\rho := \int_D u\partial_x\rho\,\mathrm{d}x$, we have

$$\begin{aligned}
\int_D u\partial_x\varphi\,\mathrm{d}x &= \int_D u\partial_{xx}\psi\,\mathrm{d}x + c_\varphi C_\rho = \int_D v\psi\,\mathrm{d}x + c_\varphi C_\rho\\
&= \int_D v(x)\int_{-1}^x (\varphi(y) - c_\varphi\rho(y))\,\mathrm{d}y\,\mathrm{d}x + c_\varphi C_\rho\\
&= \int_D \left(\int_y^1 v(x)\,\mathrm{d}x\right)(\varphi(y) - c_\varphi\rho(y))\,\mathrm{d}y + c_\varphi C_\rho.
\end{aligned}$$

Setting $C'_\rho := C_\rho - \int_D \left(\int_y^1 v(x)\,\mathrm{d}x\right)\rho(y)\,\mathrm{d}y$, we thus have

$$\int_D u\partial_x\varphi\,\mathrm{d}x = \int_D \left(\int_y^1 v(x)\,\mathrm{d}x\right)\varphi(y)\,\mathrm{d}y + C'_\rho\int_D \varphi(y)\,\mathrm{d}y,$$

which shows that u has a weak first-order derivative.

Exercise 2.5 (Clairaut's theorem). Let $\varphi \in C_0^\infty(D)$. Using Clairaut's theorem for φ, we infer that

$$\begin{aligned}(-1)^{|\alpha|}\int_D \partial^\alpha v\partial^\beta\varphi\,\mathrm{d}x &= \int_D v\partial^\alpha(\partial^\beta\varphi)\,\mathrm{d}x\\ &= \int_D v\partial^\beta(\partial^\alpha\varphi)\,\mathrm{d}x = (-1)^{|\beta|}\int_D \partial^\beta v\partial^\alpha\varphi\,\mathrm{d}x,\end{aligned}$$

where we used the definition of $\partial^\alpha v$ (and $\partial^\beta\varphi \in C_0^\infty(D)$) and $\partial^\beta v$ (and $\partial^\alpha\varphi \in C_0^\infty(D)$). Using the definition of $\partial^\alpha(\partial^\beta v)$, the above identity shows that

$$\int_D \partial^\alpha v\partial^\beta\varphi\,\mathrm{d}x = (-1)^{|\beta|}\int_D \partial^\alpha(\partial^\beta v)\varphi\,\mathrm{d}x,$$

for all $\varphi \in C_0^\infty(D)$, which, in turn, implies that the weak derivative $\partial^\beta(\partial^\alpha v)$ exists and that this weak derivative is indeed equal to $\partial^\alpha(\partial^\beta v)$.

Exercise 2.6 (Weak and classical derivatives). Let $\alpha \in \mathbb{N}^d$ be a multi-index of length $|\alpha| \le k$. Let $(\partial^\alpha v)_{\mathrm{cl}}$, $(\partial^\alpha v)_{\mathrm{wk}}$ denote the classical and weak derivatives, respectively. For all $\varphi \in C_0^\infty(D)$, integrating by parts the classical derivative (there are no boundary terms since φ has compact support), we infer that

$$\int_D (\partial^\alpha v)_{\mathrm{cl}}\varphi\,\mathrm{d}x = (-1)^{|\alpha|}\int_D v\partial^\alpha\varphi\,\mathrm{d}x = \int_D (\partial^\alpha v)_{\mathrm{wk}}\varphi\,\mathrm{d}x,$$

and we conclude by invoking the vanishing integral theorem (Theorem 1.32).

Exercise 2.7 ($H^1(D)$). (i) Let us set $D := (-1,1)$. We have $u \in L^2(D)$. Let us determine whether u has a weak derivative and whether the weak derivative is in $L^2(D)$. Let $\phi \in C_0^\infty(D)$. We observe that

$$\begin{aligned}\int_{-1}^1 u(x)\partial_x\phi(x)\,\mathrm{d}x &= \int_{-1}^0((-x)^{\frac{3}{2}}-1)\partial_x\phi(x)\,\mathrm{d}x + \int_0^1 (x^{\frac{3}{2}}-1)\partial_x\phi(x)\,\mathrm{d}x\\ &= -\int_{-1}^0 -\tfrac{3}{2}(-x)^{\frac{1}{2}}\phi(x)\,\mathrm{d}x - \phi(0) - \int_0^1 \tfrac{3}{2}x^{\frac{1}{2}}\phi(x)\,\mathrm{d}x + \phi(0)\\ &= -\int_{-1}^0 -\tfrac{3}{2}|x|^{\frac{1}{2}}\phi(x)\,\mathrm{d}x - \int_0^1 \tfrac{3}{2}|x|^{\frac{1}{2}}\phi(x)\,\mathrm{d}x\\ &= -\int_{-1}^1 w(x)\phi(x)\,\mathrm{d}x,\end{aligned}$$

where $w(x) := \frac{3}{2}|x|^{\frac{1}{2}}\mathrm{sgn(x)}$ with

$$\mathrm{sgn}(x) = \begin{cases}-1 & \text{if } x<0,\\ 1 & \text{otherwise.}\end{cases}$$

Since $w \in L^2(D)$, we infer that $u \in H^1(D)$.
(ii) Since $u_1 \in L^2((-1,0))$, $u_2 \in L^2((0,1))$ and $\|u\|_{L^2(D)} = (\|u_1\|^2_{L^2((-1,0))} + \|u_2\|^2_{L^2((0,1))})^{\frac{1}{2}}$, we infer that $u \in L^2(D)$. Let $\phi \in C_0^\infty(D)$. Using that $u_1 \in C^1((-1,0])$, $u_2 \in C^1([0,1))$, and setting $v(x) := u_1(x)$ if $x < 0$ and $v(x) := u_2(x)$ otherwise, we infer that

$$\begin{aligned}\int_D u(x)\partial_x\phi(x)\,\mathrm{d}x &= \int_{-1}^0 u_1(x)\partial_x\phi(x)\,\mathrm{d}x + \int_{-1}^0 u_2(x)\partial_x\phi(x)\,\mathrm{d}x\\ &= -\int_D v(x)\phi(x)\,\mathrm{d}x + \phi(0)(u_1(0) - u_2(0))\\ &= -\int_D v(x)\phi(x)\,\mathrm{d}x,\end{aligned}$$

since $u_1(0) = u_2(0)$. We infer that $\partial_x u = v$ which is in $L^2(D)$. Hence, $u \in H^1(D)$. If $u_1(0) \neq u_2(0)$, we infer from Example 2.5 that there is no function $w \in L^1_{\mathrm{loc}}(D)$ s.t. $\int_D v\phi\,\mathrm{d}x = \phi(0)$ for all $\phi \in C_0^\infty(D)$. Hence, $u \notin H^1(D)$.

Exercise 2.8 (Broken seminorm). (i) Let $v \in W^{1,1}_{\mathrm{loc}}(D)$. Let $k \in \{1{:}d\}$, let $i \in \{1{:}n\}$, and let $\varphi \in C_0^\infty(D_i)$. Letting $\tilde\varphi$ denote the zero-extension of φ to D, we have

$$\int_{D_i} (\partial_k v)_{|D_i}\varphi\,\mathrm{d}x = \int_D \partial_k v\tilde\varphi\,\mathrm{d}x = -\int_D v\partial_k\tilde\varphi\,\mathrm{d}x = -\int_{D_i} v_{|D_i}\partial_k\varphi\,\mathrm{d}x,$$

which shows that $(\partial_k v)_{|D_i} = \partial_k(v_{|D_i})$.
(ii) The identity is a direct consequence of $(\nabla v)_{|D_i} = \nabla(v_{|D_i})$ since

$$\begin{aligned}\sum_{i\in\{1:n\}} |v_{|D_i}|^p_{W^{1,p}(D_i)} &= \sum_{i\in\{1:n\}} \|\nabla(v_{|D_i})\|^p_{\boldsymbol{L}^p(D_i)} = \sum_{i\in\{1:n\}} \|(\nabla v)_{|D_i}\|^p_{\boldsymbol{L}^p(D_i)}\\ &= \|\nabla v\|^p_{\boldsymbol{L}^p(D)} = |v|^p_{W^{1,p}(D)}.\end{aligned}$$

(iii) The definition of $W^{s,p}(D)$ implies that

$$\begin{aligned}|v|^p_{W^{s,p}(D)} &= \int_D\int_D \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^p}{\|\boldsymbol{x}-\boldsymbol{y}\|^{sp+d}_{\ell^2}}\,\mathrm{d}x\,\mathrm{d}y\\ &= \sum_{i\in\{1:n\}}\sum_{j\in\{1:n\}} \int_{D_i}\int_{D_j} \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^p}{\|\boldsymbol{x}-\boldsymbol{y}\|^{sp+d}_{\ell^2}}\,\mathrm{d}x\,\mathrm{d}y\\ &\geq \sum_{i\in\{1:n\}} \int_{D_i}\int_{D_i} \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^p}{\|\boldsymbol{x}-\boldsymbol{y}\|^{sp+d}_{\ell^2}}\,\mathrm{d}x\,\mathrm{d}y = \sum_{i\in\{1:n\}} |v|^p_{W^{s,p}(D_i)}.\end{aligned}$$

Exercise 2.9 ($W^{s,p}$). Notice first that $C^{0,\alpha}(D) \hookrightarrow L^\infty(D)$ since $\alpha > 0$. Then $C^{0,\alpha}(D) \hookrightarrow L^p(D)$ since D is bounded. Let $v \in C^{0,\alpha}(D)$ and let c_α be the constant such that $|v(\boldsymbol{x}) - v(\boldsymbol{y})| \leq c_\alpha\|\boldsymbol{x}-\boldsymbol{y}\|^\alpha_{\ell^2}$. Let $\ell_D := \mathrm{diam}(D)$. Since $D \subset B(\boldsymbol{x}, \ell_D)$ for all $\boldsymbol{x} \in D$, we infer that

$$\begin{aligned}|v|^p_{W^{s,p}(D)} &= \int_D\int_D \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^p}{\|\boldsymbol{x}-\boldsymbol{y}\|^{sp+d}_{\ell^2}}\,\mathrm{d}x\,\mathrm{d}y \leq c_\alpha^p\int_D\int_D \frac{1}{\|\boldsymbol{x}-\boldsymbol{y}\|^{(s-\alpha)p+d}_{\ell^2}}\,\mathrm{d}x\,\mathrm{d}y\\ &\leq c_\alpha^p\int_D\int_{B(\boldsymbol{x},\ell_D)} \frac{1}{\|\boldsymbol{x}-\boldsymbol{y}\|^{(s-\alpha)p+d}_{\ell^2}}\,\mathrm{d}x\,\mathrm{d}y\\ &= c_\alpha^p|S(\boldsymbol{0},1)|\ell_D^{(\alpha-s)p}\int_D\int_0^1 r^{(\alpha-s)p-1}\,\mathrm{d}r\,\mathrm{d}x,\end{aligned}$$

where $S(\mathbf{0},1)$ is the unit sphere in $\mathbb{R}^d$ and $|S(\mathbf{0},1)|$ is the $(d-1)$-dimensional measure of $S(\mathbf{0},1)$. The integral is finite, i.e., $|v|_{W^{s,p}(D)}$ is finite, if and only if $s<\alpha$.

Exercise 2.10 (Unbounded function in $H^1(D)$). (i) First, we observe that $u\in L^2(D)$. Let $\epsilon\in(0,\frac12)$. Then u is of class C^∞ in $D\backslash B(\mathbf{0},\epsilon)$. Using radial coordinates, we set $w_1(\boldsymbol{x}):=\partial_1 u(\boldsymbol{x})=\frac{\cos(\theta)}{r\ln(r)}$ and $w_2(\boldsymbol{x}):=\partial_2 u(\boldsymbol{x})=\frac{\sin(\theta)}{r\ln(r)}$ for $\boldsymbol{x}\neq\mathbf{0}$. One can verify that $w_i\in L^1(D)$ for $i\in\{1,2\}$. Let $\varphi\in C_0^\infty(D)$. We obtain

$$\int_{D\backslash B(\mathbf{0},\epsilon)} w_i\varphi\,\mathrm{d}x=-\int_{D\backslash B(\mathbf{0},\epsilon)} u\partial_i\varphi\,\mathrm{d}x+T(\epsilon),$$

with $T(\epsilon):=\int_{\partial B(\mathbf{0},\epsilon)}(\boldsymbol{e}_i\cdot\boldsymbol{n})u\varphi\,\mathrm{d}s$, where $\boldsymbol{n}$ is the unit normal at $\partial B(\mathbf{0},\epsilon)$ pointing outward and $\boldsymbol{e}_i$ is the unit canonical vector defining the i-th direction. Since $|T(\epsilon)|\le 2\pi\epsilon\ln(-\ln(\epsilon))\|\varphi\|_{L^\infty(D)}$, we infer that $T(\epsilon)\to 0$ as $\epsilon\to 0$. Since w_i and u are in $L^1(D)$, letting $\epsilon\to 0$ in the above equality, we infer using Lebesgue's dominated convergence theorem that $\int_D w_i\varphi\,\mathrm{d}x=-\int_D u\partial_i\varphi\,\mathrm{d}x$. Since φ is arbitrary in $C_0^\infty(D)$, we conclude that w_i is the weak derivative of u in the i-th direction.
(ii) Let us now show that $w_i\in L^2(D)$ for all $i\in\{1,2\}$. We have

$$\|w_1\|_{L^2(D)}^2+\|w_2\|_{L^2(D)}^2=2\pi\int_0^{\frac12}\frac{r}{r^2\ln(r)^2}\,\mathrm{d}r=2\pi\int_{-\infty}^{-\ln(2)}\frac{1}{t^2}\,\mathrm{d}t=\frac{2\pi}{\ln(2)}<\infty.$$

Exercise 2.11 (Equivalent norm). For every integer $n\ge 0$, let $\mathcal{B}_{n,d}:=\{\alpha\in\mathbb{N}^d\mid|\alpha|=n\}$ and $b_{n,d}:=\mathrm{card}(\mathcal{B}_{n,d})$. Let $Y_n:=[L^p(D)]^{b_{n,d}}$ be equipped with some product norm. Let us consider the integer $m\ge 2$ and let $Y:=L^p(D)\times Y_m$ be equipped with the product norm

$$\|(z,(y_\alpha)_{\alpha\in\mathcal{B}_{m,d}})\|_Y:=\left(\|z\|_{L^p(D)}^p+\ell_D^{mp}\sum_{\alpha\in\mathcal{B}_{m,d}}\|y_\alpha\|_{L^p(D)}^p\right)^{\frac1p}.$$

We define the operator $A:W^{m,p}(D)\to Y$ by $A(v):=(v,(\partial^\alpha v)_{\alpha\in\mathcal{B}_{m,d}})$. Let $Z:=Y_1\times\ldots\times Y_{m-1}$ be equipped with some product norm. We define $T:W^{m,p}(D)\to Z$ by

$$T(v):=((\partial^{\alpha_1}v)_{\alpha_1\in\mathcal{B}_{1,d}},\ldots,(\partial^{\alpha_{m-1}}v)_{\alpha_{m-1}\in\mathcal{B}_{m-1,d}}).$$

The equivalence of norms in finite-dimensional spaces implies that there exists c such that

$$c\,\|v\|_{W^{m,p}(D)}\le\|A(v)\|_Y+\|T(v)\|_Z,\qquad\forall v\in W^{m,p}(D).$$

Both A and T are linear and bounded. The operator A is injective. The operator T is compact since the embedding $W^{m,p}(D)\hookrightarrow W^{m',p}(D)$ is compact for all $m'\in\{1{:}m{-}1\}$ owing to the compact embeddings from Theorem 2.35. Then the assertion follows from Lemma A.20.

Chapter 3

Traces and Poincaré inequalities

Exercises

Exercise 3.1 (Scaling). Let $D \subset \mathbb{R}^d$ be a Lipschitz domain. Let $\lambda > 0$ and $\tilde{D} := \lambda^{-1} D$. (i) Show that D and $\tilde{D}$ have the same Poincaré–Steklov constant in (3.8). (ii) Same question for (3.11).

Exercise 3.2 (Poincaré–Steklov, 1D). Let $D := (0,1)$ and $u \in C^1(D;\mathbb{R})$. Prove the following bounds: (i) $\|u\|_{L^2(D)}^2 \le \frac{1}{2}\|u'\|_{L^2(D)}^2$ if $u(0) = 0$. (*Hint*: $u(x) = \int_0^x u'(t)\,\mathrm{d}t$.) (ii) $\|u\|_{L^2(D)}^2 \le \frac{1}{\sqrt{8}}\|u'\|_{L^2(D)}^2$ if $u(0) = u(1) = 0$. (*Hint*: as above, but distinguish whether $x \in (0, \frac{1}{2})$ or $x \in (\frac{1}{2}, 1)$.) (iii) $\|u\|_{L^2(D)}^2 \le \frac{1}{6}\|u'\|_{L^2(D)}^2 + \underline{u}^2$ with $\underline{u} := \int_0^1 u\,\mathrm{d}x$. (*Hint*: square the identity $u(x) - u(y) = \int_x^y u'(t)\,\mathrm{d}t$.) (iv) $\max_{x\in\overline{D}} |u(x)|^2 \le 2u(1)^2 + 2\|u'\|_{L^2(D)}^2$. (*Hint*: square $u(x) = u(1) + \int_1^x u'(t)\,\mathrm{d}t$.) (v) $\max_{x\in\overline{D}} |u(x)|^2 \le 2(\|u\|_{L^2(D)}^2 + \|u'\|_{L^2(D)}^2)$. (*Hint*: prove that $u(x)^2 \le 2u(y)^2 + 2\|u'\|_{L^2(D)}^2$ and integrate over $y \in D$.)

Exercise 3.3 (Fractional Poincaré–Steklov). (i) Prove (3.10). (*Hint*: write $\int_D |v(\boldsymbol{x}) - \underline{v}_D|^p \mathrm{d}x = \int_D |D|^{-p} \left| \int_D (v(\boldsymbol{x}) - v(\boldsymbol{y}))\,\mathrm{d}y \right|^p \mathrm{d}x$.) (ii) Prove that $|v - \underline{v}_D|_{W^{r,p}(D)} \le \ell_D^{s-r} |v|_{W^{s,p}(D)}$ for all $r \in (0, s]$ and all $s \in (0,1)$.

Exercise 3.4 (Zero-extension in $W_0^{1,p}(D)$). Let $p \in [1,\infty)$. Let D be an open set in $\mathbb{R}^d$. Show that $W_0^{1,p}(D) \hookrightarrow \widetilde{W}^{1,p}(D)$ and $\|\tilde{u}\|_{W^{1,p}(\mathbb{R}^d)} \le \|u\|_{W^{1,p}(D)}$ for all $u \in W_0^{1,p}(D)$.

Exercise 3.5 (Integral representation). Let $v : [0,\infty) \longrightarrow \mathbb{R}$ be a continuous function with bounded derivative, and let $w : [0,\infty) \longrightarrow \mathbb{R}$ be such that $w(x) := \frac{1}{x}\int_0^x (v(t) - v(x))\,\mathrm{d}t$. (i) Show that $|w(x)| \le \frac{Mx}{2}$ where $M := \sup_{x\in[0,\infty)} |\partial_x v(x)|$. (ii) Estimate $w(0)$. (iii) Show that $\partial_t(tw(t)) = -t\partial_t v(t)$. (iv) Prove that $v(x) - v(0) = -w(x) - \int_0^x \frac{w(t)}{t}\,\mathrm{d}t$. (*Hint*: observe that $v(x) - v(0) = \int_0^x \frac{1}{t}\,(t\partial_t v(t))\,\mathrm{d}t$, use (iii), and integrate by parts.) (v) Prove the following integral representation formula (see Grisvard [20, pp. 29-30]):

$$v(0) = v(x) + \frac{1}{x}\int_0^x (v(t) - v(x))\,\mathrm{d}t + \int_0^x \frac{1}{y^2}\int_0^y (v(t) - v(y))\,\mathrm{d}t\,\mathrm{d}y.$$

Exercise 3.6 (Trace inequality in $W^{s,p}$, $sp > 1$). Let $s \in (0,1)$, $p \in [1,\infty)$, and $sp > 1$. Let $a > 0$ and F be an open bounded subset of $\mathbb{R}^{d-1}$. Let $D := F \times (0,a)$. Let $v \in C^1(D) \cap C^0(\overline{D})$. (i)

Let $\boldsymbol{y} \in F$. Using the integral representation from Exercise 3.5, show that there are $c_1(s,p)$ and $c_2(s,p)$ such that

$$|v(\boldsymbol{y},0)| \le a^{-\frac{1}{p}}\|v(\boldsymbol{y},.)\|_{L^p(0,a)} + (c_1(s,p)+c_2(s,p))a^{s-\frac{1}{p}}|v(\boldsymbol{y},.)|_{W^{s,p}(0,a)}.$$

(ii) Accept as a fact that there is c (depending on s and p) such that

$$\int_F \int_0^a \int_0^a \frac{|v(\boldsymbol{x}_{d-1},x_d)-v(\boldsymbol{x}_{d-1},y_d)|^p}{|x_d-y_d|^{sp+1}}\,\mathrm{d}x_1\ldots\,\mathrm{d}x_{d-1}\,\mathrm{d}x_d\,\mathrm{d}y_d \le c\,|v|_{W^{s,p}(D)}.$$

Prove that $\|v(.,0)\|_{L^p(F)} \le c'\,(a^{-\frac{1}{p}}\|v\|_{L^p(D)} + a^{s-\frac{1}{p}}|v|_{W^{s,p}(D)})$. *Note*: this shows that the trace operator $\gamma^{\mathrm{g}} : C^1(D)\cap C^0(\overline{D}) \to L^p(F)$ is bounded uniformly w.r.t. the norm of $W^{s,p}(D)$ when $sp>1$. This means that γ^{g} can be extended to $W^{s,p}(D)$ since $C^1(D)\cap C^0(\overline{D})$ is dense in $W^{s,p}(D)$.

Solution to exercises

Exercise 3.1 (Scaling). (i) Consider the mapping $\boldsymbol{\psi} : \tilde{D} \to D$ s.t. $\boldsymbol{\psi}(\widetilde{\boldsymbol{x}}) := \lambda\widetilde{\boldsymbol{x}}$. We have $\boldsymbol{\psi}^{-1}(\boldsymbol{x}) = \lambda^{-1}\boldsymbol{x}$. Let $\mathbb{J}$ be the Jacobian matrix of $\boldsymbol{\psi}$, i.e., $\mathbb{J} = \lambda\mathbb{I}_d$ and $\mathbb{J}^{-1} = \lambda^{-1}\mathbb{I}_d$ (where $\mathbb{I}_d$ is the identity matrix in $\mathbb{R}^{d\times d}$). Let $v \in W^{1,p}(D)$ and set $\tilde{v}(\widetilde{\boldsymbol{x}}) := v(\boldsymbol{\psi}(\widetilde{\boldsymbol{x}}))$ for all $\widetilde{\boldsymbol{x}} \in \tilde{D}$. We infer that $(\nabla\tilde{v})(\widetilde{\boldsymbol{x}}) = \lambda(\nabla v)(\boldsymbol{\psi}(\widetilde{\boldsymbol{x}}))$ and

$$\|\nabla\tilde{v}\|^p_{\boldsymbol{L}^p(\tilde{D})} = \lambda^p\int_D \|\nabla v(\boldsymbol{x})\|^p|\det(\mathbb{J}^{-1})|\,\mathrm{d}x = \lambda^{p-d}\int_D\|\nabla v(\boldsymbol{x})\|^p\,\mathrm{d}x.$$

Hence, $\|\nabla\tilde{v}\|_{\boldsymbol{L}^p(\tilde{D})} = \lambda^{1-\frac{d}{p}}\|\nabla v\|_{\boldsymbol{L}^p(D)}$. Moreover, using that $|\tilde{D}| = \lambda^{-d}|D|$, we infer that

$$\underline{\tilde{v}}_{\tilde{D}} = \frac{1}{|\tilde{D}|}\int_{\tilde{D}}\tilde{v}\,\mathrm{d}\tilde{x} = \frac{\lambda^d}{|D|}\int_D v|\det(\mathbb{J}^{-1})|\,\mathrm{d}x = \underline{v}_D.$$

Hence, we have

$$\|\tilde{v}-\underline{\tilde{v}}_{\tilde{D}}\|^p_{L^p(\tilde{D})} = \int_D (v-\underline{v}_D)^p|\det(\mathbb{J}^{-1})|\,\mathrm{d}x = \lambda^{-d}\|v-\underline{v}_D\|^p_{L^p(D)}.$$

Assume that (3.8) holds true for all $v \in W^{1,p}(D)$. Using that $\ell_{\tilde{D}} = \lambda^{-1}\ell_D$, we obtain

$$\begin{aligned} C_{\mathrm{PS},p}\|\tilde{v}-\underline{\tilde{v}}_{\tilde{D}}\|_{L^p(\tilde{D})} &= \lambda^{-\frac{d}{p}}C_{\mathrm{PS},p}\|v-\underline{v}_D\|_{L^p(D)} \\ &\le \lambda^{-\frac{d}{p}}\ell_D|v|_{W^{1,p}(D)} \\ &= \lambda^{1-\frac{d}{p}}\ell_{\tilde{D}}\lambda^{-1+\frac{d}{p}}|\tilde{v}|_{W^{1,p}(\tilde{D})} \\ &= \ell_{\tilde{D}}|\tilde{v}|_{W^{1,p}(\tilde{D})}, \end{aligned}$$

which proves the assertion.
(ii) The proof for (3.11) is similar.

Exercise 3.2 (Poincaré–Steklov, 1D). Let $u \in C^1(D;\mathbb{R})$.
(i) Assume that that $u(0)=0$. Then we have

$$\begin{aligned} |u(x)| &= \left|\int_0^x u'(t)\,\mathrm{d}t\right| \le \int_0^x |u'(t)|\,\mathrm{d}t \\ &\le \left(\int_0^x \mathrm{d}t\right)^{\frac{1}{2}}\left(\int_0^x (u'(t))^2\,\mathrm{d}t\right)^{\frac{1}{2}} \le x^{\frac{1}{2}}\|u'\|_{L^2(D)}. \end{aligned}$$

This implies that $\|u\|^2_{L^2(D)} \le (\int_0^1 x\,\mathrm{d}x)\|u'\|^2_{L^2(D)} = \frac{1}{2}\|u'\|^2_{L^2(D)}$ if $u(0) = 0$.
(ii) Assume that $u(0) = u(1) = 0$. If $x \in (0, \frac{1}{2})$, the above argument shows that $|u(x)| \le x^{\frac{1}{2}}\|u'\|_{L^2(0,\frac{1}{2})}$. Similarly, if $x \in (\frac{1}{2}, 1)$, we have $|u(x)| \le (1-x)^{\frac{1}{2}}\|u'\|_{L^2(\frac{1}{2},1)}$. We infer that

$$\begin{aligned}\|u\|^2_{L^2(D)} &= \int_0^{\frac{1}{2}} u(x)^2\,\mathrm{d}x + \int_{\frac{1}{2}}^1 u(x)^2\,\mathrm{d}x\\ &\le \|u'\|^2_{L^2(0,\frac{1}{2})}\int_0^{\frac{1}{2}} x\,\mathrm{d}x + \|u'\|^2_{L^2(\frac{1}{2},1)}\int_{\frac{1}{2}}^1 (1-x)\,\mathrm{d}x\\ &\le \frac{1}{8}\Big(\|u'\|^2_{L^2(0,\frac{1}{2})} + \|u'\|^2_{L^2(\frac{1}{2},1)}\Big) = \frac{1}{8}\|u'\|^2_{L^2(0,1)}.\end{aligned}$$

(iii) Let us set $\underline{u} := \int_0^1 u\,\mathrm{d}x$. After squaring the equation $u(x) - u(y) = \int_x^y u'(t)\,\mathrm{d}t$, we obtain

$$u(x)^2 + u(y)^2 - 2u(y)u(x) = \int_x^y u'(t)\,\mathrm{d}t \le |y-x|\,\|u'\|^2_{L^2(D)}.$$

This, in turn, implies that

$$\int_D\int_D u(x)^2\,\mathrm{d}x\,\mathrm{d}y + \int_D\int_D u(y)^2\,\mathrm{d}x\,\mathrm{d}y - 2\int_D\int_D u(y)u(x)\,\mathrm{d}x\,\mathrm{d}y \le \|u'\|^2_{L^2(D)}\int_D\int_D |y-x|\,\mathrm{d}x\,\mathrm{d}y.$$

A direct computation shows that

$$\int_D\int_D u(x)^2\,\mathrm{d}x\,\mathrm{d}y + \int_D\int_D u(y)^2\,\mathrm{d}x\,\mathrm{d}y - 2\int_D\int_D u(y)u(x)\,\mathrm{d}x\,\mathrm{d}y = 2\|u\|^2_{L^2(D)} - 2\underline{u}^2,$$

and that

$$\begin{aligned}\int_D\int_D |y-x|\,\mathrm{d}x\,\mathrm{d}y &= \int_0^1\left(\int_0^y (y-x)\,\mathrm{d}x + \int_y^1 (x-y)\,\mathrm{d}x\right)\mathrm{d}y\\ &= \int_0^1 (\frac{1}{2}y^2 + \frac{1}{2}(1-y)^2)\,\mathrm{d}y = \frac{2}{6} = \frac{1}{3}.\end{aligned}$$

We conclude that

$$2\|u\|^2_{L^2(D)} - 2\underline{u}^2 \le \frac{1}{3}\|u'\|^2_{L^2(D)},$$

which proves the assertion.
(iv) Let $x \in \overline{D} = [0,1]$. Recalling that $u(x) = u(1) + \int_1^x u'(t)\,\mathrm{d}t$, we obtain

$$\begin{aligned}u(x)^2 &= \left(u(1) + \int_1^x u'(t)\,\mathrm{d}t\right)^2 \le 2u(1)^2 + 2\left(\int_1^x u'(t)\,\mathrm{d}t\right)^2\\ &\le 2u(1)^2 + 2x\|u'\|^2_{L^2(D)} \le 2u(1)^2 + 2\|u'\|^2_{L^2(D)},\end{aligned}$$

which proves that $\max_{x\in\overline{D}} |u(x)|^2 \le 2u^2(1) + 2\|u'\|^2_{L^2(D)}$.
(v) Similarly, we have $u(x) = u(y) + \int_y^x u'(t)\,\mathrm{d}t$. Proceeding as above, we obtain

$$\begin{aligned}u(x)^2 &= \left(u(y) + \int_y^x u'(t)\,\mathrm{d}t\right)^2 \le 2u(y)^2 + 2\left(\int_y^x u'(t)\,\mathrm{d}t\right)^2\\ &\le 2u(y)^2 + 2|y-x|\|u'\|^2_{L^2(D)} \le 2u(y)^2 + 2\|u'\|^2_{L^2(D)}.\end{aligned}$$

This implies that

$$u(x)^2 \int_D \mathrm{d}y \le 2 \int_D u(y)^2 \,\mathrm{d}y + 2\|u'\|_{L^2(D)}^2 \int_D \mathrm{d}y,$$

and we conclude that $\max_{x\in\overline{D}} |u(x)|^2 \le 2\|u\|_{L^2(D)}^2 + 2\|u'\|_{L^2(D)}^2$.

Exercise 3.3 (Fractional Poincaré–Steklov). (i) Following the hint, we observe that

$$\begin{aligned}
\int_D |v(\boldsymbol{x}) - \underline{v}_D|^p \mathrm{d}x &= \int_D |D|^{-p} \left| \int_D (v(\boldsymbol{x}) - v(\boldsymbol{y}))\,\mathrm{d}y \right|^p \mathrm{d}x \\
&\le \int_D |D|^{-p} \left(\int_D \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|}{\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2}^{s+\frac{d}{p}}} \|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2}^{s+\frac{d}{p}} \,\mathrm{d}y \right)^p \mathrm{d}x \\
&\le \int_D |D|^{-p} \int_D \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^p}{\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2}^{sp+d}} \,\mathrm{d}y \left(\int_D \|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2}^{(s+\frac{d}{p})p'} \,\mathrm{d}y \right)^{\frac{p}{p'}} \mathrm{d}x,
\end{aligned}$$

where $p' := \frac{p}{p-1}$. Using that $\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2} \le \ell_D$ for all $\boldsymbol{x},\boldsymbol{y} \in D$, we infer that

$$\begin{aligned}
\|v - \underline{v}_D\|_{L^p(D)}^p &\le \int_D |D|^{-p} \int_D \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^p}{\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2}^{sp+d}} \,\mathrm{d}y\,\mathrm{d}x \left(\max_{\boldsymbol{x}\in D} \int_D \|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2}^{(s+\frac{d}{p})p'} \,\mathrm{d}y \right)^{\frac{p}{p'}} \\
&\le |v|_{W^{s,p}(D)}^p |D|^{-p} \left(\int_D \ell_D^{(s+\frac{d}{p})p'} \,\mathrm{d}y \right)^{\frac{p}{p'}} \\
&\le |v|_{W^{s,p}(D)}^p |D|^{-p} |D|^{\frac{p}{p'}} \ell_D^{sp+d} \le |v|_{W^{s,p}(D)}^p \ell_D^{sp+d} |D|^{-1}.
\end{aligned}$$

Hence, $\|v - \underline{v}_D\|_{L^p(D)} \le \ell_D^s \left(\frac{\ell_D^d}{|D|}\right)^{\frac{1}{p}} |v|_{W^{s,p}(D)}$.
(ii) Using the definitions and $\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2} \le \ell_D$ for all $\boldsymbol{x},\boldsymbol{y} \in D$, we have

$$\begin{aligned}
|v - \underline{v}_D|_{W^{r,p}(D)}^p = |v|_{W^{r,p}(D)}^p &= \int_D \int_D \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^p}{\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2}^{rp+d}} \,\mathrm{d}x\,\mathrm{d}y \\
&= \int_D \int_D \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^p}{\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2}^{sp+d}} \|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2}^{(s-r)p} \,\mathrm{d}x\,\mathrm{d}y \\
&\le \ell_D^{(s-r)p} |v|_{W^{s,p}(D)}^p.
\end{aligned}$$

This concludes the proof.

Exercise 3.4 (Zero-extension in $W_0^{1,p}(D)$). Let $u \in W_0^{1,p}(D)$. By definition, there is a sequence $(u_n)_{n\in\mathbb{N}}$ in $C_0^\infty(D)$ such that $u_n \to u$ in $W^{1,p}(D)$. For all $\varphi \in C_0^\infty(\mathbb{R}^d)$ and all $i \in \{1{:}d\}$, we observe that

$$\begin{aligned}
\int_{\mathbb{R}^d} \tilde{u} \partial_i \varphi \,\mathrm{d}x = \int_D u \partial_i \varphi \,\mathrm{d}x &= \lim_{n\to\infty} \int_D u_n \partial_i \varphi \,\mathrm{d}x = -\lim_{n\to\infty} \int_D \partial_i u_n \varphi \,\mathrm{d}x \\
&= \int_D \partial_i u \varphi \,\mathrm{d}x \le \|\partial_i u\|_{L^p(D)} \|\varphi\|_{L^{p'}(D)} \\
&\le \|\partial_i u\|_{L^p(D)} \|\varphi\|_{L^{p'}(\mathbb{R}^d)},
\end{aligned}$$

where $p' \in (1,\infty]$ is the conjugate of p. This shows that the linear form $\varphi \mapsto \int_{\mathbb{R}^d} \tilde{u}\partial_i\varphi \,\mathrm{d}x$ is bounded in $L^{p'}(\mathbb{R}^d)$. Since $L^{p'}(\mathbb{R}^d) = (L^p(\mathbb{R}^d))'$, we infer that $\partial_i\tilde{u} \in L^p(\mathbb{R}^d)$. Since $i \in \{1{:}d\}$ is arbitrary, this implies that $\tilde{u} \in W^{1,p}(\mathbb{R}^d)$, i.e., $u \in \widetilde{W}^{1,p}(D)$. Finally, the estimate $\|\tilde{u}\|_{W^{1,p}(\mathbb{R}^d)} \le \|u\|_{W^{1,p}(D)}$ results from the above bound.

Exercise 3.5 (Integral representation). (i) Any time we see a quantity like $v(t) - v(x)$, we must think of the fundamental theorem of calculus, i.e., $v(t) - v(x) = \int_x^t \partial_x v(z)\,\mathrm{d}z$. We have

$$\begin{aligned}|w(x)| &= \frac{1}{x}\left|\int_0^x (v(t)-v(x))\,\mathrm{d}t\right| = \frac{1}{x}\left|\int_0^x\int_x^t \partial_z v(z)\,\mathrm{d}z\,\mathrm{d}t\right| \\ &\le \frac{1}{x}\int_0^x\left|\int_x^t \partial_z v(z)\,\mathrm{d}z\right|\mathrm{d}t \le \frac{1}{x}\int_0^x\int_t^x |\partial_z v(z)|\,\mathrm{d}z\,\mathrm{d}t \\ &\le \frac{M}{x}\int_0^x\int_t^x \mathrm{d}z\,\mathrm{d}t = \frac{M}{x}\int_0^x (x-t)\,\mathrm{d}t = \frac{M}{x}(x^2 - \frac{1}{2}x^2).\end{aligned}$$

Hence, $|w(x)| \le \frac{Mx}{2}$ for all $x \in [0,\infty)$.
(ii) The estimate $|w(x)| \le \frac{Mx}{2}$ shows that $|w(0)| \le 0$, meaning that $w(0) = 0$.
(iii) Upon observing that $tw(t) = \int_0^t (v(z) - v(t))\,\mathrm{d}z$ and recalling that the fundamental theorem of calculus implies that

$$\partial_t\left(\int_0^t f(z)\,\mathrm{d}z\right) = f(t),$$

we have

$$\begin{aligned}\partial(tw(t)) &= \partial_t\int_0^t (v(z)-v(t))\,\mathrm{d}z = \partial_t\int_0^t v(z)\,\mathrm{d}z - \partial_t(v(t)t) \\ &= v(t) - v(t) - t\partial_t v(t) = -t\partial_t v(t).\end{aligned}$$

Hence, $\partial_t(tw(t)) = -t\partial_t v(t)$.
(iv) Following the hint, we infer that

$$\begin{aligned}v(x) - v(0) &= \int_0^x \frac{1}{t}\left(t\partial_t v(t)\right)\,\mathrm{d}t = -\int_0^x \frac{1}{t}\partial_t(tw(t))\,\mathrm{d}t \\ &= \int_0^x \partial_t\left(\frac{1}{t}\right) tw(t)\,\mathrm{d}t - \left[\frac{1}{t}tw(t)\right]_0^x \\ &= -\int_0^x \frac{1}{t^2}tw(t)\,\mathrm{d}t - w(x) + w(0),\end{aligned}$$

thereby proving that $v(x) - v(0) = -\int_0^x \frac{1}{t}w(t)\,\mathrm{d}t - w(x)$.
(v) The integral representation is obtained by replacing $w(t)$ and $w(x)$ in the above identity.

Exercise 3.6 (Trace inequality in $W^{s,p}$, $sp > 1$). The identity from Exercise 3.5 gives

$$v(\boldsymbol{y},0) = v(\boldsymbol{y},x) + \frac{1}{x}\int_0^x (v(\boldsymbol{y},t) - v(\boldsymbol{y},x))\,\mathrm{d}t + \int_0^x \frac{1}{y^2}\int_0^y (v(\boldsymbol{y},t) - v(\boldsymbol{y},y))\,\mathrm{d}t\,\mathrm{d}y.$$

Using Hölder's inequality repeatedly, we infer that

$$\frac{1}{a}\int_0^a v(\boldsymbol{y},x)\,\mathrm{d}x \le a^{-\frac{1}{p}}\|v(\boldsymbol{y},.)\|_{L^p(0,a)},$$
$$\frac{1}{a}\int_0^a \frac{1}{x}\int_0^x (v(\boldsymbol{y},t)-v(\boldsymbol{y},x))\,\mathrm{d}t\,\mathrm{d}x \le c_1(s,p)\,a^{s-\frac{1}{p}}|v(\boldsymbol{y},.)|_{W^{s.p}(0,a)},$$
$$\frac{1}{a}\int_0^a \int_0^x \frac{1}{y^2}\int_0^y (v(\boldsymbol{y},t)-v(\boldsymbol{y},y))\,\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}x \le c_2(s,p)\,a^{s-\frac{1}{p}}|v(\boldsymbol{y},.)|_{W^{s.p}(0,a)},$$

where $c_1(s,p) := \left(\frac{p-1}{p(s+1)}\frac{p-1}{p(s+1)-1}\right)^{\frac{p-1}{p}}$, $c_2(s,p) := \left(\frac{p-1}{p(s+1)}\frac{p-1}{sp-1}\right)^{\frac{p-1}{p}}\frac{p}{p(s+1)-1}$. Using that $v(\boldsymbol{y},0) = \frac{1}{a}\int_0^a v(\boldsymbol{y},0)\,\mathrm{d}x$, we infer that

$$|v(\boldsymbol{y},0)| \le a^{-\frac{1}{p}}\|v(\boldsymbol{y},.)\|_{L^p(0,a)} + (c_1(s,p)+c_2(s,p))a^{s-\frac{1}{p}}|v(\boldsymbol{y},.)|_{W^{s,p}(0,a)}.$$

(ii) Using the inequality $(\alpha+\beta)^p \le 2^{\frac{p-1}{p}}(|\alpha|^p+|\beta|^p)$, we infer that

$$\|v(.,0)\|_{L^p(F)} \le c\,(a^{-\frac{1}{p}}\|v\|_{L^p(D)} + a^{s-\frac{1}{p}}I(v)),$$

where

$$I(v)^p := \int_F\int_0^a\int_0^a \frac{|v(\boldsymbol{x}_{d-1},x_d)-v(\boldsymbol{x}_{d-1},y_d)|^p}{|x_d-y_d|^{sp+1}}\,\mathrm{d}x_1\ldots\,\mathrm{d}x_{d-1}\,\mathrm{d}x_d\,\mathrm{d}y_d.$$

The rest of the proof consists of proving that there is a constant c such that $I(v) \le c|v|_{W^{s,p}(F\times(0,a))}$. This is actually (a slightly modified version of) Lemma 4.33 in [13, p. 200].

Chapter 4

Distributions and duality in Sobolev spaces

Exercises

Exercise 4.1 (Distributions). Let D be an open set in $\mathbb{R}^d$. Let v be a distribution in D. (i) Let $\psi \in C^\infty(D)$. Show that the map $C_0^\infty(D) \ni \varphi \mapsto \langle v, \psi\varphi\rangle$ defines a distribution in D (this distribution is usually denoted by ψv). (ii) Let $\alpha, \beta \in \mathbb{N}^d$. Prove that $\partial^\alpha(\partial^\beta v) = \partial^\beta(\partial^\alpha v)$ in the distribution sense.

Exercise 4.2 (Dirac measure on a manifold). Let D be a smooth bounded and open set in $\mathbb{R}^d$. Let $u \in C^2(D;\mathbb{R})$ and assume that $u_{|\partial D} = 0$. Let $\tilde{u}$ be the extension by zero of u over $\mathbb{R}^d$. Compute $\nabla\cdot(\nabla\tilde{u}) = \partial_{11}u + \ldots + \partial_{dd}u$ in the distribution sense.

Exercise 4.3 (P.V. $\frac{1}{x}$). Let $D := (-1,1)$. Prove that the linear map $T : C_0^\infty(D) \to \mathbb{R}$ defined by $\langle T, \varphi\rangle := \lim_{\epsilon\to 0} \int_{|x|>|\epsilon|} \frac{1}{x}\varphi(x)\,\mathrm{d}x$ is a distribution.

Exercise 4.4 (Integration by parts). Prove the two identities in (4.8) by using the divergence formula $\int_D \nabla\cdot\boldsymbol{\phi}\,\mathrm{d}x = \int_{\partial D}(\boldsymbol{\phi}\cdot\boldsymbol{n})\,\mathrm{d}s$ for all $\boldsymbol{\phi} \in \boldsymbol{C}^1(\overline{D})$.

Exercise 4.5 (Definition (4.11)). Verify that the right-hand side of (4.11) is independent of the choice of $\boldsymbol{w}(\boldsymbol{l})$. (*Hint*: consider two functions $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \boldsymbol{W}^{1,p'}(D)$ s.t. $\gamma^{\mathrm{g}}(\boldsymbol{w}_1) = \gamma^{\mathrm{g}}(\boldsymbol{w}_2) = \boldsymbol{l}$ and use the density of $\boldsymbol{C}_0^\infty(D)$ in $\boldsymbol{W}_0^{1,p'}(D)$.)

Solution to exercises

Exercise 4.1 (Distributions). (i) Let K be a compact subset of the open set D. Since v is a distribution, there exist $c \in \mathbb{R}$ and $p \in \mathbb{N}$ (both depending on K) such that $|\langle v, \varphi\rangle| \le c \max_{|\alpha|\le p}\left(\ell_D^{|\alpha|}\|\partial^\alpha\varphi\|_{L^\infty(K)}\right)$ for all $\varphi \in C_0^\infty(D)$ (i.e., φ is a smooth function with compact support in K). Let $\psi \in C^\infty(D)$ and $\varphi \in C_0^\infty(D)$. Since $\psi\varphi \in C^p(D)$ and $\mathrm{supp}(\psi\varphi) \subset K$, we have $\psi\varphi \in \mathcal{C}_0^p(D)$. The product rule implies that there exists c' depending on p such that $\|\partial^\alpha(\psi\varphi)\|_{L^\infty(K)} \le$

$c' \max_{|\beta|\le|\alpha|} \big(\ell_D^{|\beta|}\|\partial^\beta\psi\|_{L^\infty(K)}\big)\|\partial^\alpha\varphi\|_{L^\infty(K)}$ for all α s.t. $|\alpha|\le p$. For all $\varphi\in C_0^\infty(D)$, we infer that

$$\begin{aligned}|\langle v,\psi\varphi\rangle| &\le c\max_{|\alpha|\le p}\big(\ell_D^{|\alpha|}\|\partial^\alpha(\psi\varphi)\|_{L^\infty(K)}\big)\\ &\le c\,c'\max_{|\alpha|\le p}\max_{|\beta|\le|\alpha|}\big(\ell_D^{|\beta|}\|\partial^\beta\psi\|_{L^\infty(K)}\big)\|\partial^\alpha\varphi\|_{L^\infty(K)}\\ &\le c\,c'\max_{|\beta|\le p}\big(\ell_D^{|\beta|}\|\partial^\beta\psi\|_{L^\infty(K)}\big)\max_{|\alpha|\le p}\|\partial^\alpha\varphi\|_{L^\infty(K)},\end{aligned}$$

thereby proving that the linear map $C_0^\infty(D)\ni\varphi\mapsto\langle v,\psi\varphi\rangle\in\mathbb{R}$ is a distribution.
(ii) Let $\alpha,\beta\in\mathbb{N}^d$. Let $\varphi\in C_0^\infty(D)$. Using Clairaut's theorem for φ, we infer that

$$\begin{aligned}\langle\partial^\beta(\partial^\alpha v),\varphi\rangle &= (-1)^{|\beta|}\langle\partial^\alpha v,(\partial^\beta\varphi)\rangle = (-1)^{|\alpha|+|\beta|}\langle v,\partial^\alpha(\partial^\beta\varphi)\rangle\\ &= (-1)^{|\alpha|+|\beta|}\langle v,\partial^\beta(\partial^\alpha\varphi)\rangle = (-1)^{|\alpha|}\langle\partial^\beta v,(\partial^\alpha\varphi)\rangle\\ &= \langle\partial^\alpha(\partial^\beta v),\varphi\rangle.\end{aligned}$$

Hence, $\partial^\beta(\partial^\alpha v)=\partial^\alpha(\partial^\beta v)$, which is Clairaut's theorem for distributions.

Exercise 4.2 (Dirac measure on a manifold). We use the notation $\Delta:=\nabla\cdot(\nabla)$. By definition, we have

$$\begin{aligned}\langle\Delta\tilde u,\varphi\rangle &= \langle\tilde u,\partial_{11}\varphi+\ldots+\partial_{dd}\varphi\rangle = \int_{\mathbb{R}^d}\tilde u\nabla\cdot(\nabla\varphi)\,\mathrm{d}x = \int_D u\nabla\cdot(\nabla\varphi)\,\mathrm{d}x\\ &= -\int_D\nabla u\cdot\nabla\varphi\,\mathrm{d}x = \int_D\nabla\cdot(\nabla u)\varphi\,\mathrm{d}x - \int_{\partial D}(\boldsymbol{n}\cdot\nabla u)\varphi\,\mathrm{d}s\\ &= \int_{\mathbb{R}^d}\widetilde{\Delta u}\varphi\,\mathrm{d}x - \int_{\partial D}(\boldsymbol{n}\cdot\nabla u)\varphi\,\mathrm{d}s.\end{aligned}$$

Hence, we have proved that $\Delta\tilde u=\widetilde{\Delta u}-(\nabla u\cdot\boldsymbol{n})\delta_{\partial D}$, where $\delta_{\partial D}$ is the Dirac measure whose support is ∂D. *Note*: one can make sense of the notation $(\boldsymbol{n}\cdot\nabla u)\delta_{\partial D}$ by smoothly extending $\boldsymbol{n}$ and u over $\mathbb{R}^d$ and by reasoning as in Exercise 4.1(i) with $p:=0$.

Exercise 4.3 (P.V. $\frac{1}{x}$). Let $\epsilon>0$. We have

$$\begin{aligned}\int_{|x|>|\epsilon|}\frac{1}{x}\varphi(x)\,\mathrm{d}x &= \int_{-1}^{-\epsilon}\frac{1}{x}\varphi(x)\,\mathrm{d}x+\int_\epsilon^1\frac{1}{x}\varphi(x)\,\mathrm{d}x\\ &= -\int_{-1}^{-\epsilon}\ln(|x|)\varphi'(x)\,\mathrm{d}x+\varphi(-\epsilon)\ln(\epsilon)-\int_\epsilon^1\ln(x)\varphi'(x)\,\mathrm{d}x-\varphi(\epsilon)\ln(\epsilon)\\ &= \int_{-1}^1\mathbb{1}_{|(-1,-\epsilon)\cup(\epsilon,1)}\ln(|x|)\varphi'(x)\,\mathrm{d}x+(\varphi(-\epsilon)-\varphi(\epsilon))\ln(\epsilon),\end{aligned}$$

where $\mathbb{1}_E$ is the indicator function of the set E. We notice that

$$|(\varphi(-\epsilon)-\varphi(\epsilon))\ln(\epsilon)|\le\|\varphi'\|_{L^\infty}\epsilon\ln(\epsilon).$$

Moreover, the sequence $\mathbb{1}_{|(-1,-\epsilon)\cup(\epsilon,1)}\ln(|x|)\varphi'(x)$ converges a.e. in D to $\ln(|x|)\varphi'(x)$, and we also have $\mathbb{1}_{|(-1,-\epsilon)\cup(\epsilon,1)}\ln(|x|)\varphi'(x)\le\ln(|x|)\varphi'(x)\in L^1(D)$. Lebesgue's dominated convergence implies that

$$\langle T,\varphi\rangle:=\lim_{\epsilon\to0}\int_{|x|>|\epsilon|}\frac{1}{x}\varphi(x)\,\mathrm{d}x=-\int_{-1}^1\ln(|x|)\varphi'(x)\,\mathrm{d}x,$$

i.e., the limit process with respect to ϵ is well defined. Moreover, we have

$$|\langle T, \varphi\rangle| = \left| \int_{-1}^{1} x(\ln(|x|) - 1)\varphi''(x)\,\mathrm{d}x \right| \leq \|\varphi''(x)\|_{L^\infty(D)},$$

thereby proving that T is indeed a distribution. Notice that we have actually proved that $T = \partial_x(\frac{1}{x})$, which makes sense after all.

Exercise 4.4 (Integration by parts). The identity (4.8a) follows from the divergence formula for $\boldsymbol{\phi}$ by using $\boldsymbol{\phi} := \boldsymbol{v}\times\boldsymbol{w}$ (since $\nabla\cdot(\boldsymbol{v}\times\boldsymbol{w}) = (\nabla\times\boldsymbol{v})\cdot\boldsymbol{w} - \boldsymbol{v}\cdot\nabla\times\boldsymbol{w}$) and $\boldsymbol{\phi}\cdot\boldsymbol{n} = (\boldsymbol{v}\times\boldsymbol{w})\times\boldsymbol{n} = -(\boldsymbol{v}\times\boldsymbol{n})\cdot\boldsymbol{w}$), whereas the identity (4.8b) follows from the divergence formula for $\boldsymbol{\phi}$ by using $\boldsymbol{\phi} := \boldsymbol{v}q$ (since $\nabla\cdot(\boldsymbol{v}q) = \boldsymbol{v}\cdot\nabla q + (\nabla\cdot\boldsymbol{v})q$ and $\boldsymbol{\phi}\cdot\boldsymbol{n} = (\boldsymbol{v}\cdot\boldsymbol{n})q$).

Exercise 4.5 (Definition (4.11)**).** Let $\boldsymbol{v} \in \boldsymbol{Z}^{\mathrm{c},p}(D)$. Following the hint, let $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \boldsymbol{W}^{1,p'}(D)$ be s.t. $\gamma^{\mathrm{g}}(\boldsymbol{w}_1) = \gamma^{\mathrm{g}}(\boldsymbol{w}_2) = \boldsymbol{l}$. Then $\boldsymbol{w}_1 - \boldsymbol{w}_2 \in \boldsymbol{W}_0^{1,p'}(D)$. Invoking a density argument, let $(\boldsymbol{\phi}_n)_{n\in\mathbb{N}}$ be a sequence in $\boldsymbol{C}_0^\infty(D)$ converging to $\boldsymbol{w}_1 - \boldsymbol{w}_2$ in $\boldsymbol{W}_0^{1,p'}(D)$. Then we have

$$0 = \int_D \boldsymbol{v}\cdot\nabla\times\boldsymbol{\phi}_n\,\mathrm{d}x - \int_D \boldsymbol{\phi}_n\cdot\nabla\times\boldsymbol{v}\,\mathrm{d}x.$$

Passing to the limit $n \to \infty$ yields

$$0 = \int_D \boldsymbol{v}\cdot\nabla\times(\boldsymbol{w}_1 - \boldsymbol{w}_2)\,\mathrm{d}x - \int_D (\boldsymbol{w}_1 - \boldsymbol{w}_2)\cdot\nabla\times\boldsymbol{v}\,\mathrm{d}x.$$

Hence, $\langle\gamma^{\mathrm{c}}(\boldsymbol{v}), \gamma^{\mathrm{g}}(\boldsymbol{w}_1)\rangle_{\partial D} = \langle\gamma^{\mathrm{c}}(\boldsymbol{v}), \gamma^{\mathrm{g}}(\boldsymbol{w}_2)\rangle_{\partial D}$, which establishes the claim.

Chapter 5

Main ideas and definitions

Exercises

Exercise 5.1 (Linear combination). Let $\mathcal{S} \in \mathbb{R}^{n_{\mathrm{sh}} \times n_{\mathrm{sh}}}$ be an invertible matrix. Let (K, P, Σ) be a finite element. Let $\tilde{\Sigma} := \{\tilde{\sigma}_i\}_{i \in \mathcal{N}}$ with dofs $\tilde{\sigma}_i := \sum_{i' \in \mathcal{N}} \mathcal{S}_{ii'} \sigma_{i'}$ for all $i \in \mathcal{N}$. Prove that $(K, P, \tilde{\Sigma})$ is a finite element. Write the shape functions $\{\tilde{\theta}_j\}_{j \in \mathcal{N}}$ and verify that the interpolation operator does not depend on $\mathcal{S}$, i.e., $\tilde{\mathcal{I}}_K(v)(\boldsymbol{x}) = \mathcal{I}_K(v)(\boldsymbol{x})$ for all $v \in V(K)$ and all $\boldsymbol{x} \in K$.

Exercise 5.2 (Modal finite element). (i) Let (K, P, Σ) and $(K, P, \widetilde{\Sigma})$ be two modal finite elements. Let $\{\zeta_i\}_{i \in \mathcal{N}}$, $\{\tilde{\zeta}_i\}_{i \in \mathcal{N}}$, be the two bases of P s.t. the dofs in Σ and $\widetilde{\Sigma}$ are given by $\sigma_i(p) := |K|^{-1}(\zeta_i, p)_{L^2(K;\mathbb{R}^q)}$ and $\tilde{\sigma}_i(p) := |K|^{-1}(\tilde{\zeta}_i, p)_{L^2(K;\mathbb{R}^q)}$ for all $i \in \mathcal{N}$. Prove that the interpolation operators $\mathcal{I}_K^{\mathrm{m}}$ and $\widetilde{\mathcal{I}}_K^{\mathrm{m}}$ are identical. (ii) Prove that $(p, \mathcal{I}_K^{\mathrm{m}}(v) - v)_{L^2(K;\mathbb{R}^q)} = 0$ for all $p \in P$. (iii) Let $\mathcal{M}$ be defined by (5.12), and let $\mathcal{M}_{ij}^{\theta} := |K|^{-1}(\theta_i, \theta_j)_{L^2(K;\mathbb{R}^q)}$ for all $i, j \in \mathcal{N}$, where $\{\theta_i\}_{i \in \mathcal{N}}$ are the shape functions associated with (K, P, Σ). Prove that $\mathcal{M}^{\theta} = \mathcal{M}^{-1}$.

Exercise 5.3 (Variation on $\mathbb{P}_2$). Let $K := [0, 1]$, $P := \mathbb{P}_2$, and $\Sigma := \{\sigma_1, \sigma_2, \sigma_3\}$ be the linear forms on P s.t. $\sigma_1(p) := p(0)$, $\sigma_2(p) := 2p(\frac{1}{2}) - p(0) - p(1)$, $\sigma_3(p) := p(1)$ for all $p \in P$. Show that (K, P, Σ) is a finite element, compute the shape functions, and indicate possible choices for $V(K)$.

Exercise 5.4 (Hermite). Let $K := [0, 1]$, $P := \mathbb{P}_3$, and $\Sigma := \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$ be the linear forms on P s.t. $\sigma_1(p) := p(0)$, $\sigma_2(p) := p'(0)$, $\sigma_3(p) := p(1)$, $\sigma_4(p) := p'(1)$ for all $p \in P$. Show that (K, P, Σ) is a finite element, compute the shape functions, and indicate possible choices for $V(K)$.

Exercise 5.5 (Powell–Sabin). Consider $K := [0, 1]$ and let P be composed of the functions that are piecewise quadratic over the intervals $[0, \frac{1}{2}] \cup [\frac{1}{2}, 1]$ and are of class C^1 over K, i.e., functions in P and their first derivatives are continuous. Let $\Sigma := \{\sigma_1, \ldots, \sigma_4\}$ be the linear forms on P s.t. $\sigma_1(p) := p(0)$, $\sigma_2(p) := p'(0)$, $\sigma_3(p) := p(1)$, $\sigma_4(p) := p'(1)$. Prove that the triple (K, P, Σ) is a finite element. Verify that the first two shape functions are

$$\theta_1(t) = \begin{cases} 1 - 2t^2 & \text{if } t \in [0, \frac{1}{2}], \\ 2(1-t)^2 & \text{if } t \in [\frac{1}{2}, 1], \end{cases} \qquad \theta_2(t) = \begin{cases} t(1 - \frac{3}{2}t) & \text{if } t \in [0, \frac{1}{2}], \\ \frac{1}{2}(1-t)^2 & \text{if } t \in [\frac{1}{2}, 1], \end{cases}$$

and compute the other two shape functions. *Note*: a two-dimensional version of this finite element on triangles has been developed in [39].

Exercise 5.6 (Lebesgue constant for Lagrange element). Prove that the Lebesgue constant $\Lambda^{\mathcal{N}}$ defined in Example 5.15 is equal to $\|\mathcal{I}_K^{\mathrm{L}}\|_{\mathcal{L}(C^0(K))}$. (*Hint*: to prove $\|\mathcal{I}_K^{\mathrm{L}}\|_{\mathcal{L}(C^0(K))} \geq \Lambda^{\mathcal{N}}$, consider functions $\{\psi_i\}_{i\in\mathcal{N}}$ taking values in $[0,1]$ s.t. $\sum_{i\in\mathcal{N}}\psi_i = 1$ in K and $\psi_i(\boldsymbol{a}_j) = \delta_{ij}$ for all $i,j\in\mathcal{N}$.)

Exercise 5.7 (Lagrange interpolation). Let $K := [a,b]$ and let $p\in[1,\infty)$. (i) Prove that $\|v\|_{L^\infty(K)} \leq (b-a)^{-\frac{1}{p}}\|v\|_{L^p(K)} + (b-a)^{1-\frac{1}{p}}\|v'\|_{L^p(K)}$ for all $v\in W^{1,p}(K)$ (*Hint*: use $v(x) - v(y) = \int_x^y v'(t)\,\mathrm{d}t$ for all $v\in C^1(K)$, where $|v(y)| := \min_{z\in K}|v(z)|$, then use the density of $C^1(K)$ in $W^{1,p}(K)$.) (ii) Prove that $W^{1,p}(K)$ embeds continuously in $C^0(K)$. (iii) Let $\mathcal{I}_K^{\mathrm{L}}$ be the interpolation operator based on the linear Lagrange finite element using the nodes a and b. Determine the two shape functions and prove that $\mathcal{I}_K^{\mathrm{L}}$ can be extended to $W^{1,p}(K)$. (iv) Assuming that $w\in W^{1,p}(K)$ is zero at some point in K, show that $\|w\|_{L^p(K)} \leq (b-a)\|w'\|_{L^p(K)}$. (v) Prove the following estimates: $\|(v-\mathcal{I}_K^{\mathrm{L}}(v))'\|_{L^p(K)} \leq (b-a)\|v''\|_{L^p(K)}$, $\|v-\mathcal{I}_K^{\mathrm{L}}(v)\|_{L^p(K)} \leq (b-a)\|(v-\mathcal{I}_K^{\mathrm{L}}(v))'\|_{L^p(K)}$, $\|(\mathcal{I}_K^{\mathrm{L}}(v))'\|_{L^p(K)} \leq \|v'\|_{L^p(K)}$, for all $p\in(1,\infty]$ and all $v\in W^{2,p}(K)$.

Exercise 5.8 (Cross approximation). Let X,Y be nonempty subsets of $\mathbb{R}$ and $f: X\times Y\to\mathbb{R}$ be a bivariate function. Let $\mathcal{N} := \{1{:}n_{\mathrm{sh}}\}$ with $n_{\mathrm{sh}}\geq 1$, and consider n_{sh} points $\{x_i\}_{i\in\mathcal{N}}$ in X and n_{sh} points $\{y_j\}_{j\in\mathcal{N}}$ in Y. Assume that the matrix $\mathcal{F}\in\mathbb{R}^{n_{\mathrm{sh}}\times n_{\mathrm{sh}}}$ with entries $\mathcal{F}_{ij} := f(x_i,y_j)$ is invertible. Let $\mathcal{I}^{\mathrm{CA}}(f): X\times Y\to\mathbb{R}$ be s.t. $\mathcal{I}^{\mathrm{CA}}(f)(x,y) := \sum_{i,j\in\mathcal{N}}(\mathcal{F}^{-\mathsf{T}})_{ij}f(x,y_j)f(x_i,y)$. Prove that $\mathcal{I}^{\mathrm{CA}}(f)(x,y_k) = f(x,y_k)$ for all $x\in X$ and all $k\in\mathcal{N}$, and that $\mathcal{I}^{\mathrm{CA}}(f)(x_k,y) = f(x_k,y)$ for all $y\in Y$ and all $k\in\mathcal{N}$.

Exercise 5.9 (Riesz–Fréchet in finite dimension). Let V be a finite-dimensional complex Hilbert space. Show that for every antilinear form $A\in V'$, there is a unique $v\in V$ s.t. $(v,w)_V = \langle A,w\rangle_{V',V}$ for all $w\in V$, with $\|v\|_V = \|A\|_{V'}$.

Solution to exercises

Exercise 5.1 (Linear combination). We use Remark 5.3. Let $p\in P$ be such that $\tilde\sigma_i(p) = 0$ for all $i\in\mathcal{N}$. The matrix $\mathcal{S}$ being invertible, we infer that $\sigma_i(p) = 0$ for all $i\in\mathcal{N}$, so that $p = 0$ since (K,P,Σ) is a finite element.
The shape functions are such that $\tilde\theta_j = \sum_{j'\in\mathcal{N}}(S^{-\mathsf{T}})_{jj'}\theta_{j'}$ since

$$\begin{aligned}\tilde\sigma_i(\tilde\theta_j) &= \sum_{i'\in\mathcal{N}}\sum_{j'\in\mathcal{N}}\mathcal{S}_{ii'}(\mathcal{S}^{-\mathsf{T}})_{jj'}\sigma_{i'}(\theta_{j'}) = \sum_{i'\in\mathcal{N}}\sum_{j'\in\mathcal{N}}\mathcal{S}_{ii'}(\mathcal{S}^{-\mathsf{T}})_{jj'}\delta_{i'j'}\\ &= \sum_{i'\in\mathcal{N}}\mathcal{S}_{ii'}(\mathcal{S}^{-\mathsf{T}})_{ji'} = \delta_{ij}.\end{aligned}$$

As a result, we infer that for all $v\in V(K)$ and all $\boldsymbol{x}\in K$,

$$\begin{aligned}\tilde{\mathcal{I}}_K(v)(\boldsymbol{x}) &= \sum_{i\in\mathcal{N}}\tilde\sigma_i(v)\tilde\theta_i(\boldsymbol{x}) = \sum_{i\in\mathcal{N}}\sum_{i'\in\mathcal{N}}\sum_{j'\in\mathcal{N}}\mathcal{S}_{ii'}(\mathcal{S}^{-\mathsf{T}})_{ij'}\sigma_{i'}(v)\theta_{j'}(\boldsymbol{x})\\ &= \sum_{i'\in\mathcal{N}}\sum_{j'\in\mathcal{N}}\delta_{i'j'}\sigma_{i'}(v)\theta_{j'}(\boldsymbol{x}) = \sum_{i'\in\mathcal{N}}\sigma_{i'}(v)\theta_{i'}(\boldsymbol{x}) = \mathcal{I}_K(v)(\boldsymbol{x}).\end{aligned}$$

Exercise 5.2 (Modal finite element). (i) Since $\{\zeta_i\}_{i\in\mathcal{N}}$ is a basis of P, there are real numbers

$\mathcal{S}_{ij}$, $i,j \in \mathcal{N}$, such that $\tilde{\zeta}_i = \sum_{j\in\mathcal{N}} \mathcal{S}_{ij}\zeta_j$. Hence, we have

$$\begin{aligned}\tilde{\sigma}_i(p) &= |K|^{-1}(\tilde{\zeta}_i, p)_{L^2(K;\mathbb{R}^q)}\\ &= \sum_{j\in\mathcal{N}} \mathcal{S}_{ij}|K|^{-1}(\zeta_j,p)_{L^2(K;\mathbb{R}^q)} = \sum_{j\in\mathcal{N}} \mathcal{S}_{ij}\sigma_j(p).\end{aligned}$$

We conclude by invoking the result from Exercise 5.1.
(ii) Let $\{\theta_i\}_{i\in\mathcal{N}}$ be the shape functions of (K,P,Σ). For all the basis functions ζ_l, we have

$$\begin{aligned}(\zeta_l, \mathcal{I}_K^{\mathrm{m}}(v),\zeta_l)_{L^2(K;\mathbb{R}^q)} &= \sum_{i\in\mathcal{N}} \sigma_i(v)(\zeta_l,\theta_i)_{L^2(K;\mathbb{R}^q)} = \sum_{i\in\mathcal{N}} \sigma_i(v)|K|\sigma_l(\theta_i)\\ &= \sigma_l(v)|K| = (\zeta_l, v)_{L^2(K;\mathbb{R}^q)}.\end{aligned}$$

This implies that $(p, \mathcal{I}_K^{\mathrm{m}}(v) - v, p)_{L^2(K;\mathbb{R}^q)} = 0$ for all $p \in P$.
(iii) Using the definitions, we have

$$\begin{aligned}(\mathcal{M}^\theta\mathcal{M})_{ij} &= \sum_{k\in\mathcal{N}} |K|^{-1}(\theta_i,\theta_k)_{L^2(K;\mathbb{R}^q)}|K|^{-1}(\zeta_k,\zeta_j)_{L^2(K;\mathbb{R}^q)}\\ &= |K|^{-1}\Big(\theta_i, \sum_{k\in\mathcal{N}} \theta_k|K|^{-1}(\zeta_k,\zeta_j)_{L^2(K;\mathbb{R}^q)}\Big)_{L^2(K;\mathbb{R}^q)}\\ &= |K|^{-1}\Big(\theta_i, \sum_{k\in\mathcal{N}} \theta_k\sigma_k(\zeta_j)\Big)_{L^2(K;\mathbb{R}^q)}\\ &= |K|^{-1}(\theta_i,\zeta_j)_{L^2(K;\mathbb{R}^q)} = \sigma_j(\theta_i) = \delta_{ij}.\end{aligned}$$

Exercise 5.3 (Variation on $\mathbb{P}_2$). Observe that $\dim\mathbb{P}_2 = 3 = \operatorname{card}\Sigma$. Let $p(x) \in \mathbb{P}_2$ be such that $\sigma_1(p) = \sigma_2(p) = \sigma_3(p) = 0$. Then $p(0) = p(1) = 0$, which implies that $0 = 2p(\frac{1}{2}) - p(0) - p(1) = 2p(\frac{1}{2})$, i.e., $p(0) = p(1) = p(\frac{1}{2}) = 0$. This, in turn, implies that p vanishes identically. One verifies that the shape functions are $\theta_1(x) = 1-x$, $\theta_2(x) = 2x(1-x)$, $\theta_3(x) = x$. Possible choices for the domain of the interpolation operator are $C^0(K)$ and $H^s(K)$ with $s > \frac{1}{2}$.

Exercise 5.4 (Hermite). We use Remark 5.3. First, we have $\dim P = \operatorname{card}\Sigma = 4$. Moreover, if $p \in \mathbb{P}_3$ is such that $\sigma_i(p) = 0$ for all $i \in \{1{:}4\}$, we infer that both t^2 and $(t-1)^2$ divide p. Since p is of degree ≤ 3, p vanishes identically.
A direct computation shows that

$$\begin{aligned}\theta_1(t) &= (2t+1)(t-1)^2, &\qquad \theta_2(t) &= t(t-1)^2,\\ \theta_3(t) &= (3-2t)t^2, &\qquad \theta_4(t) &= (t-1)t^2.\end{aligned}$$

For instance, $(t-1)^2$ divides θ_1 since $\theta_1(1) = 0$ and $\theta_1'(1) = 0$. Then $\theta_1(t) = (at+b)(t-1)^2$, and the coefficients a and b are determined by the conditions $1 = \theta_1(0) = b$ and $0 = \theta_1'(0) = a - 2b$. Note that by symmetry, we have $\theta_3(t) = \theta_1(1-t)$ and $\theta_4(t) = -\theta_2(1-t)$. Possible choices for the domain of the interpolation operator are $V(K) := C^1(K)$ or $V(K) := H^2(K)$.

Exercise 5.5 (Powell–Sabin). We use Remark 5.3. First, we have $\dim P = \operatorname{card}\Sigma = 4$. Moreover, if $p \in P$ is such that $\sigma_i(p) = 0$ for all $i \in \{1{:}4\}$, we infer that $p_{|[0,\frac{1}{2}]} = at^2$ and $p_{|[\frac{1}{2},1]} = b(1-t)^2$ for some real numbers a, b. The C^1-matching condition at $t = \frac{1}{2}$ leads to $a = b$ and $2a = -2b$, whence $a = b = 0$. By symmetry, we have $\theta_3(t) = \theta_1(1-t)$ and $\theta_4(t) = -\theta_2(1-t)$.

Exercise 5.6 (Lebesgue constant for Lagrange element). Let us prove that $\|\mathcal{I}_K^{\mathrm{L}}\|_{\mathcal{L}(C^0(K))} \le \Lambda^{\mathcal{N}}$. For all $v \in C^0(K)$, we observe that

$$|\mathcal{I}_K^{\mathrm{L}}(v)(\boldsymbol{x})| \le \sum_{i\in\mathcal{N}} |v(\boldsymbol{a}_i)||\theta_i(\boldsymbol{x})| \le \left(\sum_{i\in\mathcal{N}} |\theta_i(\boldsymbol{x})|\right) \|v\|_{C^0(K)},$$

which proves the expected bound. Let us prove the reverse bound. Using the hint, we define the function

$$v_0(\boldsymbol{x}) := \sum_{j\in\mathcal{N}} \mathrm{sgn}(\theta_j(\boldsymbol{x}_0))\psi_j(\boldsymbol{x}),$$

where $\boldsymbol{x}_0$ is a point in K where the function $\sum_{j\in\mathcal{N}} |\theta_j(\boldsymbol{x})|$ is maximal. Owing to the properties of the functions $\{\psi_i\}_{i\in\mathcal{N}}$, we infer that $\|v_0\|_{C^0(K)} = 1$, and by construction, we obtain

$$\|\mathcal{I}_K^{\mathrm{L}}(v_0)\|_{C^0(K)} \ge \mathcal{I}_K^{\mathrm{L}}(v_0)(\boldsymbol{x}_0) = \sum_{j\in\mathcal{N}} |\theta_j(\boldsymbol{x}_0)| = \Lambda^{\mathcal{N}}.$$

The functions $\{\psi_i\}_{i\in\mathcal{N}}$ can be taken to be the one-dimensional hat basis functions associated with the $\mathbb{P}_1$-Lagrange finite element. Assume that the set $\{\boldsymbol{a}_i\}_{i\in\mathcal{N}}$ contains the interval endpoints, i.e., $K = [\boldsymbol{a}_1, \boldsymbol{a}_{n_{\mathrm{sh}}}]$. Let $\psi_i : K \to [0,1]$ be the piecewise affine function s.t. $\psi_i(\boldsymbol{x}) := \frac{\boldsymbol{x}-\boldsymbol{a}_{i-1}}{\boldsymbol{a}_i-\boldsymbol{a}_{i-1}}$ if $\boldsymbol{x} \in [\boldsymbol{a}_{i-1}, \boldsymbol{a}_i]$ (and $i > 1$), $\psi_i(\boldsymbol{x}) := \frac{\boldsymbol{a}_{i+1}-\boldsymbol{x}}{\boldsymbol{a}_{i+1}-\boldsymbol{a}_i}$ if $\boldsymbol{x} \in [\boldsymbol{a}_i, \boldsymbol{a}_{i+1}]$ (and $i < n_{\mathrm{sh}}$), and $\psi(\boldsymbol{x}) := 0$ otherwise. By construction, ψ_i takes values in $[0,1]$ and $\psi_i(\boldsymbol{a}_j) = \delta_{ij}$. Moreover, the function $\sum_{i\in\mathcal{N}} \psi_i(\boldsymbol{x})$ is affine in each interval $[\boldsymbol{a}_j, \boldsymbol{a}_{j+1}]$ and takes the value 1 at the two endpoints for all $j \in \{1{:}n_{\mathrm{sh}}{-}1\}$. Hence, $\sum_{i\in\mathcal{N}} \psi_i(\boldsymbol{x}) = 1$. Finally, if $\{\boldsymbol{a}_i\}_{i\in\mathcal{N}}$ does not contain the interval endpoints, the function ψ_1 (resp., $\psi_{n_{\mathrm{sh}}}$) is extended by the constant value 1 on the left of $\boldsymbol{a}_1$ (resp., right of $\boldsymbol{a}_{n_{\mathrm{sh}}}$).

Exercise 5.7 (Lagrange interpolation). (i) Let $v \in C^1(K)$ and let $y \in K$ be such that $|v(y)| = \min_{z\in K} |v(z)|$. Since $v(x) = v(y) + \int_y^x v'(t)\,\mathrm{d}t$ for all $x \in K$, we infer using Hölder's inequality that

$$|v(x)| \le |v(y)| + (b-a)^{1-\frac{1}{p}} \|v'\|_{L^p(K)}.$$

Moreover, integrating the inequality $|v(y)| \le |v(z)|$ with respect to $z \in K$, we obtain $(b-a)^{\frac{1}{p}}|v(y)| \le \|v\|_{L^p(K)}$. We infer that

$$|v(x)| \le (b-a)^{-\frac{1}{p}} \|v\|_{L^p(K)} + (b-a)^{1-\frac{1}{p}} \|v'\|_{L^p(K)}, \qquad \forall x \in K.$$

Let now $v \in W^{1,p}(K)$. Let $(v_n)_{n\in\mathbb{N}}$ be a sequence in $C^\infty(K)$ converging to $v \in W^{1,p}(K)$. Then, up to a subsequence, $(v_n)_{n\in\mathbb{N}}$ converges to v a.e. in K, so that we can pass to the limit in the above inequality written for v_n and infer the expected bound.
(ii) Let $(v_n)_{n\in\mathbb{N}}$ be a sequence in $C^1(K)$ converging to $v \in W^{1,p}(K)$. Owing to the bound derived above, we infer that $(v_n)_{n\in\mathbb{N}}$ is a Cauchy sequence for the uniform norm. This sequence thus converges to some $\tilde{v} \in C^0(K)$. That $v = \tilde{v}$ a.e. in K results from the fact that $\int_K (v-\tilde{v})\varphi\,\mathrm{d}t = 0$ for all $\varphi \in C_0^\infty(\mathrm{int}(K))$, as can be inferred by passing to the limit in $\int_K v_n\varphi\,\mathrm{d}t$ and using the convergence of $(v_n)_{n\in\mathbb{N}}$ in $W^{1,p}(K)$ and in $C^0(K)$.
(iii) The two shape functions are $\theta_1(t) = \frac{b-t}{b-a}$ and $\theta_2(t) = \frac{t-a}{b-a}$. The extension of $\mathcal{I}_K^{\mathrm{L}}$ to $W^{1,p}(K)$ is a direct consequence of (ii).
(iv) Proceeding as in (i) with $y \in K$ such that $w(y) = 0$, one can prove that $|w(x)|^p \le (b-a)^{p-1}\|w'\|_{L^p(K)}^p$ for all $x \in K$. Integrating this inequality with respect to $x \in K$ yields the expected bound.

(v) Let $v \in W^{2,p}(K)$. The function $w := (v - \mathcal{I}_K^{\mathrm{L}}(v))'$ is in $W^{1,p}(K)$, and it vanishes at some point in K since $(v - \mathcal{I}_K^{\mathrm{L}}(v))$ is in $C^1(K)$ and vanishes at the two endpoints. Applying the bound derived in Step (iv) and observing that $w'' = v''$ since $\mathcal{I}_K^{\mathrm{L}}(v)$ is affine, we infer that $\|(v - \mathcal{I}_K^{\mathrm{L}}(v))'\|_{L^p(K)} \le (b-a)\|v''\|_{L^p(K)}$. By a similar reasoning, applying the bound derived in Step (iv) to the function $w := v - \mathcal{I}_K^{\mathrm{L}}(v)$ leads to $\|v - \mathcal{I}_K^{\mathrm{L}}(v)\|_{L^p(K)} \le (b-a)\|(v - \mathcal{I}_K^{\mathrm{L}}(v))'\|_{L^p(K)}$. Since $(\mathcal{I}_K^{\mathrm{L}}(v))' = \frac{v(b)-v(a)}{b-a}$, we finally infer that

$$\begin{aligned}\|(\mathcal{I}_K^{\mathrm{L}}(v))'\|_{L^p(K)} &\le (b-a)^{\frac{1}{p}-1}|v(b)-v(a)| \\ &\le (b-a)^{\frac{1}{p}-1}(b-a)^{1-\frac{1}{p}}\|v'\|_{L^p(K)} = \|v'\|_{L^p(K)}.\end{aligned}$$

Exercise 5.8 (Cross approximation). We only prove the first statement, the proof for the second one being similar. We observe that

$$\begin{aligned}\mathcal{I}^{\mathrm{CA}}(f)(x,y_k) &= \sum_{i,j\in\mathcal{N}} (\mathcal{F}^{-\mathsf{T}})_{ij} f(x,y_j) f(x_i,y_k) = \sum_{i,j\in\mathcal{N}} \mathcal{F}_{ik}(\mathcal{F}^{-\mathsf{T}})_{ij} f(x,y_j) \\ &= \sum_{j\in\mathcal{N}} \delta_{jk} f(x,y_j) = f(x,y_k).\end{aligned}$$

Exercise 5.9 (Riesz–Fréchet in finite dimension). Let $m := \dim(V)$. Let $K := \ker(A)$. The rank nullity theorem implies that $\dim(K) = m-1$. Hence, $K^\perp$ is one-dimensional, i.e., there exists $q \in V$ s.t. $K^\perp = \operatorname{span}\{q\}$ and $\|q\|_V = 1$. For all $w \in V$, we have $w = (w,q)_V q + k$ with $k \in K^\perp$. We infer that

$$\begin{aligned}\langle A, w\rangle_{V',V} &= \langle A, (w,q)_V q\rangle_{V',V} + \langle A, k\rangle_{V',V} \\ &= \overline{(w,q)_V}\langle A, q\rangle_{V',V} = (q,w)_V \langle A, q\rangle_{V',V} = ((\langle A, q\rangle_{V',V} q), w)_V.\end{aligned}$$

Denoting $v := \langle A, q\rangle_{V',V} q$, we have thus shown that $\langle A, w\rangle_{V',V} = (v,w)_V$ for all $w \in V$. The equality of norms follows from

$$\|v\|_V = \sup_{w\in V} \frac{|(v,w)_V|}{\|w\|_V} = \sup_{w\in V} \frac{|\langle A, w\rangle_{V',V}|}{\|w\|_V} = \|A\|_{V'}.$$

Finally, the uniqueness of $v \in V$ is established by contradiction. If there were distinct $v_1, v_2 \in V$ s.t. $(v_1,w)_V = \langle A, w\rangle_{V',V} = (v_2,w)_V$ for all $w \in V$, we would have $(v_1 - v_2, w)_V = 0$, and considering $w := v_1 - v_2$ leads to the expected contradiction.

Chapter 6

One-dimensional finite elements and tensorization

Exercises

Exercise 6.1 (Integrated Legendre polynomials). Let $k \geq 2$ and set $\mathbb{P}_k^{(0)} := \{p \in \mathbb{P}_k \mid p(\pm 1) = 0\}$. Show that a basis for $\mathbb{P}_k^{(0)}$ are the integrated Legendre polynomials $\{\int_{-1}^t L_l(s)\,\mathrm{d}s\}_{l\in\{1:k-1\}}$. Prove (6.6). (*Hint*: consider moments against polynomials in $\mathbb{P}_{m-2}$ and the derivative at $t = 1$.)

Exercise 6.2 (Gauss–Lobatto). The goal of this exercise is to prove Proposition 6.6. (i) Prove that $k_{\mathcal{Q}} = 2m-3$. (*Hint*: for all $p \in \mathbb{P}_{2m-3}$, $m \geq 3$, write $p = p_1(1-t^2)L'_{m-1} + p_2$ with $p_1 \in \mathbb{P}_{m-3}$ and $p_2 \in \mathbb{P}_{m-1}$.) (ii) Prove that $\omega_1 = \omega_m = \frac{2}{m(m-1)}$. (*Hint*: compute $\int_{-1}^1 L'_{m-1}(t)(1+t)L'_{m-1}(t)\,\mathrm{d}t$ using the quadrature and by integrating by parts.) (iii) Assume $m \geq 3$ and let $l \in \{2{:}m-1\}$. Prove that $L'_{m-2}(\xi_l) = (1-m)L_{m-1}(\xi_l)$ and $(1-\xi_l^2)L''_{m-1}(\xi_l)+m(m-1)L_{m-1}(\xi_l) = 0$. (*Hint*: use (6.3).) Let $\mathcal{L}_l \in \mathbb{P}_{m-3}$ be the Lagrange interpolation polynomial s.t. $\mathcal{L}_l(\xi_j) = \delta_{lj}$, for all $l, j \in \{2{:}m-1\}$ (i.e., ξ_1 and ξ_m are excluded). Prove that $\mathcal{L}_l(t) = \frac{L'_{m-1}(t)}{t-\xi_l}\frac{1}{L''_{m-1}(\xi_l)}$. (*Hint*: compare the degree of the polynomials, their roots, and their value at ξ_l.) Finally, prove (6.11). (*Hint*: integrate the polynomial $\mathcal{L}_l(t)(1-t)L'_{m-2}(t)$.) (iv) Let $\|p\|_\xi^2 := \sum_{l\in\{1:m\}} \omega_l p(\xi_l)^2$. Verify that $\|\cdot\|_\xi$ defines a norm on $\mathbb{P}_k$ with $k := m-1$, and prove that $\|p\|_{L^2(K)} \leq \|p\|_\xi \leq (\frac{2k+1}{k})^{\frac{1}{2}}\|p\|_{L^2(K)}$ for all $p \in \mathbb{P}_k$, with $K := (-1, 1)$. (*Hint*: write $p = p_{k-1} + \lambda L_k$ with $p_{k-1} \in \mathbb{P}_{k-1}$ and $\lambda \in \mathbb{R}$, and compute $\|p\|_{L^2(K)}^2$ and $\|p\|_\xi^2$.)

Exercise 6.3 (Gauss–Radau). The goal is to prove Proposition 6.7. (i) Prove that $k_{\mathcal{Q}} = 2m-2$. (*Hint*: for all $p \in \mathbb{P}_{2m-2}$, write $p = p_1(L_m - L_{m-1}) + p_2$ with $p_1 \in \mathbb{P}_{m-2}$ and $p_2 \in \mathbb{P}_{m-1}$.) (ii) Prove that $\omega_m = \frac{2}{m^2}$. (*Hint*: integrate the polynomial $\frac{L_m(t)-L_{m-1}(t)}{t-1}L'_{m-1}(t)$.) (iii) Assume $m \geq 2$ and let $l \in \{1{:}m-1\}$. Prove that $L'_m(\xi_l) = -L'_{m-1}(\xi_l)$. (*Hint*: use (6.3a) and (6.3b).) Let $\mathcal{L}_l \in \mathbb{P}_{m-2}$ be the Lagrange interpolation polynomial s.t. $\mathcal{L}_l(\xi_j) = \delta_{lj}$ for all $l, j \in \{1{:}m-1\}$ (i.e., ξ_m is excluded). Prove that $\mathcal{L}_l(t) = \frac{L_m(t)-L_{m-1}(t)}{(1-t)(t-\xi_l)}\frac{1-\xi_l}{-2L'_{m-1}(\xi_l)}$. (*Hint*: compare the degree of the polynomials, their roots, and their value at ξ_l.) Finally prove (6.12). (*Hint*: integrate the polynomial $\mathcal{L}_l(t)(1-t)L'_{m-1}(t)$.)

Exercise 6.4 (Inverse trace inequality). Let $K := [-1, 1]^d$. Let $m \geq 3$ and let $\{\xi_l\}_{l\in\{1:m\}}$ be the Gauss–Lobatto (GL) nodes in $[-1, 1]$. Set $I_{m,d} := \{1 \dots m\}^d$ and $I^0_{m,d} := \{2{:}(m-1)\}^d$. For

any $\alpha \in I_{m,d}$, let $\boldsymbol{a}_\alpha \in K$ be the node with Cartesian coordinates $(a_\alpha)_i := \xi_{\alpha_i}$ for all $i \in \{1{:}d\}$. The set $(\boldsymbol{a}_\alpha)_{\alpha\in I_{m,d}}$ consists of the tensorized GL nodes in K. Let $k := m-1$ and define the polynomial space $\mathbb{Q}^0_{k,d} := \{q \in \mathbb{Q}_{k,d} \mid q(\boldsymbol{a}_\alpha) = 0,\ \forall \alpha \in I^0_{m,d}\}$, i.e., polynomials in $\mathbb{Q}^0_{k,d}$ vanish at all the tensorized GL nodes that are located inside K. Prove that

$$\|v\|_{L^2(K)} \le \left(\frac{2d}{k(k+1)}(2+\tfrac{1}{k})^{d-1}\frac{|K|}{|\partial K|}\right)^{\frac12} \|v\|_{L^2(\partial K)},$$

for all $v \in \mathbb{Q}^0_{k,d}$. (*Hint*: use Exercise 6.2.)

Exercise 6.5 (Lagrange mass matrix). Let $\mathcal{M} \in \mathbb{R}^{n_{\mathrm{sh}}\times n_{\mathrm{sh}}}$ be the mass matrix with entries $\mathcal{M}_{ij} := \int_{-1}^1 \mathcal{L}^{[a]}_{i-1}(t)\mathcal{L}^{[a]}_{j-1}(t)\,\mathrm{d}t$ for all $i,j \in \mathcal{N}$. Prove that $\mathcal{M} = (\mathcal{V}^\mathsf{T}\mathcal{V})^{-1}$, where $\mathcal{V} \in \mathbb{R}^{n_{\mathrm{sh}}\times n_{\mathrm{sh}}}$ is the (generalized) Vandermonde matrix with entries $\mathcal{V}_{ij} := (\frac{2i-1}{2})^{\frac12} L_{i-1}(a_j)$. (*Hint*: see Proposition 5.5.)

Exercise 6.6 (Canonical hybrid element). Prove Proposition 6.10. (*Hint*: use Remark 5.3.) Compute the shape functions when $\mu_l := J^{1,1}_{l-1}$ for all $l \in \{1{:}k{-}1\}$. (*Hint*: consider the polynomials $J^{1,0}_{k-1}$, $J^{1,1}_{l-1}$ for all $l \in \{1{:}k{-}1\}$, and $J^{0,1}_{k-1}$.)

Exercise 6.7 ($\mathbb{Q}_{k,d}$ Lagrange). Prove Proposition 6.14. (*Hint*: observe that any polynomial $q \in \mathbb{Q}_{k,d}$ is such that $q(\boldsymbol{x}) = \sum_{i_d\in\{0:k\}} q_{i_d}(x_1,\dots,x_{d-1})x_d^{i_d}$ and use induction on d.)

Exercise 6.8 (Bicubic Hermite). Let K be a rectangle with vertices $\{\boldsymbol{z}_i\}_{1\le i\le 4}$, $P := \mathbb{Q}_{3,2}$, and $\Sigma := \{p(\boldsymbol{z}_i), \partial_{x_1}p(\boldsymbol{z}_i), \partial_{x_2}p(\boldsymbol{z}_i), \partial^2_{x_1x_2}p(\boldsymbol{z}_i)\}_{1\le i\le 4}$. Show that (K,P,Σ) is a finite element. (*Hint*: write $p \in \mathbb{Q}_{3,2}$ in the form $p(\boldsymbol{x}) = \sum_{i,j\in\{1:4\}} \gamma_{ij}\theta_i(x_1)\theta_j(x_2)$, where $\{\theta_1,\dots,\theta_4\}$ are the shape functions of the one-dimensional Hermite finite element; see Exercise 5.4.)

Exercise 6.9 (Face unisolvence). Prove Lemma 6.15. (*Hint*: use the hint from Exercise 6.7.)

Solution to exercises

Exercise 6.1 (Integrated Legendre polynomials). The space $\mathbb{P}^{(0)}_k$ has dimension $k-1$. Set $\theta_l(t) := \int_{-1}^t L_l(t')\,\mathrm{d}t'$ for all $l \in \{1{:}k{-}1\}$. These functions are in $\mathbb{P}^{(0)}_k$ since $\theta_l(-1) = 0$ by construction and $\theta_l(1) = 0$ by the orthogonality property of Legendre polynomials. It remains to show that the functions $\{\theta_l(t)\}_{l\in\{1:k-1\}}$ are linearly independent. Assume that $\sum_{l\in\{1:k-1\}}\alpha_l\theta_l$ vanishes identically. Taking the derivative, we infer that $\sum_{l\in\{1:k-1\}}\alpha_l L_l$ vanishes identically, which implies $\alpha_l = 0$ for all $l \in \{1{:}k{-}1\}$.
Since both sides of (6.6) are polynomials of order $(m{+}1)$ vanishing at ±1, it is enough to prove that both polynomials have the same moments against polynomials in $\mathbb{P}_{m-2}$ and that their derivative at $t = 1$ coincides. For all $\mu \in \mathbb{P}_{m-2}$, we observe that

$$-\frac{1}{2m}\int_{-1}^1 (1-t^2)J^{1,1}_{m-1}(t)\mu(t)\,\mathrm{d}t = 0,$$

and integration by parts leads to

$$\int_{-1}^1 \left(\int_{-1}^t L_m(s)\,\mathrm{d}s\right)\mu(t)\,\mathrm{d}t = -\int_{-1}^1 L_m(t)\left(\int_{-1}^t \mu(s)\,\mathrm{d}s\right)\mathrm{d}s = 0,$$

since the integrated Legendre polynomial vanishes at ±1 and since $\int_{-1}^{t}\mu(s)\,\mathrm{d}s$ is in $\mathbb{P}_{m-1}$. Furthermore, considering the derivative at $t=1$, that of the left-hand side of (6.6) is $L_m(1)=1$, whereas that of the right-hand side is $-\frac{1}{2m}(-2)J_{m-1}^{1,1}(1)=1$. This completes the proof.

Exercise 6.2 (Gauss–Lobatto). (i) We already know from Lemma 6.4 that $m-1\le k_{\mathcal{Q}}\le 2m-1$. If $m=2$, then $m-1=2m-3$ and $k_{\mathcal{Q}}\ge 2m-3$. If $m\ge 3$, let $p\in\mathbb{P}_{2m-3}$ and using the Euclidean polynomial division, write $p=p_1(1-t^2)L'_{m-1}+p_2$ with $p_1\in\mathbb{P}_{m-3}$ and $p_2\in\mathbb{P}_{m-1}$. Integrating by parts, we infer that

$$\int_{-1}^{1}p_1(t)(1-t^2)L'_{m-1}(t)\,\mathrm{d}t=-\int_{-1}^{1}\big(p_1(t)(1-t^2)\big)'\,L_{m-1}(t)\,\mathrm{d}t=0,$$

since $\big(p_1(t)(1-t^2)\big)'$ is in $\mathbb{P}_{m-2}$. Therefore, we obtain

$$\int_{-1}^{1}p(t)\,\mathrm{d}t=\int_{-1}^{1}p_2(t)\,\mathrm{d}t=\sum_{l\in\{1:m\}}\omega_l p_2(\xi_l)=\sum_{l\in\{1:m\}}\omega_l p(\xi_l),$$

where we used that $p_2(\xi_l)=p(\xi_l)$ for all $l\in\{1{:}m\}$. Hence, $k_{\mathcal{Q}}\ge 2m-3$ for all $m\ge3$ as well. For all $m\ge2$, the quadrature is not of higher order since it does not integrate exactly the polynomial $(1-t^2)(L'_{m-1})^2$ which is of degree $(2m-2)$ (the quadrature approximates its integral by zero).
(ii) Following the hint, we observe that the polynomial $L'_{m-1}(t)(1+t)L'_{m-1}(t)$ is of degree $(2m-3)$, so that it is integrated exactly by the quadrature. Since this polynomial is nonzero only at $\xi_m=1$, we infer that

$$\int_{-1}^{1}L'_{m-1}(t)(1+t)L'_{m-1}(t)\,\mathrm{d}t=\omega_m 2L'_{m-1}(1)^2=\omega_m\frac{m^2(m-1)^2}{2}.$$

Moreover, integrating by parts leads to

$$\int_{-1}^{1}L'_{m-1}(t)(1+t)L'_{m-1}(t)\,\mathrm{d}t=2L'_{m-1}(1)L_{m-1}(1)=m(m-1),$$

since the polynomial $\big(L'_{m-1}(t)(1+t)\big)'$ is of degree $(m-2)$. Combining the above two identities proves that $\omega_m=\frac{2}{m(m-1)}$. The proof that $\omega_1=\frac{2}{m(m-1)}$ is similar and consists of using the polynomial $L'_{m-1}(t)(1-t)L'_{m-1}(t)$ (one can also invoke a symmetry argument).
(iii) Let $m\ge3$ and $l\in\{2{:}m-1\}$. Applying (6.3a) at $t=\xi_l$ with the index $m-1$ yields $\xi_l L_{m-1}(\xi_l)=L_{m-2}(\xi_l)$ since $L'_{m-1}(\xi_l)=0$. Applying (6.3b) leads to $(m-1)L_{m-2}(\xi_l)+\xi_l L'_{m-2}(\xi_l)=0$. Combining these two equalities, we infer that $L'_{m-2}(\xi_l)=(1-m)L_{m-1}(\xi_l)$. Applying (6.3c) at $t=\xi_l$ with the index $m-1$ finally yields $(1-\xi_l^2)L''_{m-1}(\xi_l)+m(m-1)L_{m-1}(\xi_l)=0$ since $L'_{m-1}(\xi_l)=0$.

Let us now consider the two polynomials $\mathcal{L}_l(t)$ and $\frac{L'_{m-1}(t)}{t-\xi_l}\frac{1}{L''_{m-1}(\xi_l)}$. These polynomials are of degree $(m-3)$, they vanish at the $(m-3)$ interior Gauss–Lobatto nodes except at ξ_l where they take the common value 1. Hence, these two polynomials coincide identically.

Let us finally prove (6.11). Since the polynomial $\mathcal{L}_l(t)(1-t)L'_{m-2}(t)$ is of degree $(2m-5)$, it is integrated exactly by the quadrature. Using the quadrature and the identity $\mathcal{L}_l(t)=\frac{L'_{m-1}(t)}{t-\xi_l}\frac{1}{L''_{m-1}(\xi_l)}$, we infer that

$$\begin{aligned}\int_{-1}^{1}\mathcal{L}_l(t)(1-t)L'_{m-2}(t)\,\mathrm{d}t&=\omega_l(1-\xi_l)L'_{m-2}(\xi_l)+\frac{4\mathcal{L}_l(-1)L'_{m-2}(-1)}{m(m-1)}\\&=\omega_l(1-\xi_l)L'_{m-2}(\xi_l)+\frac{(m-1)(m-2)}{(1+\xi_l)L''_{m-1}(\xi_l)},\end{aligned}$$

where we used that $\mathcal{L}_l(-1) = -\frac{L'_{m-1}(-1)}{1+\xi_l}\frac{1}{L''_{m-1}(\xi_l)}$, together with the fact that $L'_{m-1}(-1)L'_{m-2}(-1) = -\frac{1}{4}m(m-1)^2(m-2)$. Moreover, since the polynomial $\big(\mathcal{L}_l(t)(1-t)\big)'$ is of degree $(m-3)$, integrating by parts and using the L^2-orthogonality property of L_{m-2} leads to

$$\begin{aligned}\int_{-1}^1 \mathcal{L}_l(t)(1-t)L'_{m-2}(t)\,\mathrm{d}t &= -2\mathcal{L}_l(-1)L_{m-2}(-1)\\ &= \frac{m(m-1)}{(1+\xi_l)L''_{m-1}(\xi_l)}.\end{aligned}$$

Combining the above two equalities leads to

$$\omega_l(1-\xi_l)L'_{m-2}(\xi_l) = \frac{2(m-1)}{(1+\xi_l)L''_{m-1}(\xi_l)}.$$

We conclude using the identities $L'_{m-2}(\xi_l) = -(m-1)L_{m-1}(\xi_l)$ and $(1-\xi_l^2)L''_{m-1}(\xi_l) = -m(m-1)L_{m-1}(\xi_l)$.

(iv) Let $k := m-1$. To prove that $\|\cdot\|_\xi$ defines a norm on $\mathbb{P}_k$, we need to prove that $(p,q)_\xi := \sum_{l\in\{1:m\}}\omega_l p(\xi_l)q(\xi_l)$ is an inner product on $\mathbb{P}_k$. The only nontrivial property is definiteness. Assume that $\|p\|_\xi = 0$. Since all the weights are positive, we have $p(\xi_l) = 0$ for all $l \in \{1{:}m\}$. Hence, $p = 0$ since p vanishes at $m = k+1$ distinct points and $p \in \mathbb{P}_k$. Let now $p \in \mathbb{P}_k$. Following the hint, let us write $p = p_{k-1} + \lambda L_k$ with $p_{k-1} \in \mathbb{P}_{k-1}$ and $\lambda \in \mathbb{R}$. The L^2-orthogonality of Legendre polynomials implies that

$$\|p\|^2_{L^2(K)} = \int_{-1}^1 (p_{k-1}(t) + \lambda L_k(t))^2\,\mathrm{d}t = \int_{-1}^1 p_{k-1}(t)^2\,\mathrm{d}t + \lambda^2\frac{2}{2k+1}.$$

Since p_{k-1}^2 is of degree $2k-2 = 2m-4$ and the quadrature is of order $2m-3$, we infer that $\int_{-1}^1 p_{k-1}(t)^2\,\mathrm{d}t = \|p_{k-1}\|_\xi^2$. Moreover, owing to (6.11), we infer that $\|L_k\|_\xi^2 = \frac{2}{k}$. Hence, we have

$$\|p\|^2_{L^2(K)} = \|p_{k-1}\|_\xi^2 + \frac{k}{2k+1}\lambda^2\|L_k\|_\xi^2.$$

In addition, since the polynomial $p_{k-1}(t)L_k(t)$ is of degree $2k-1 = 2m-3$, we infer that $0 = \int_{-1}^1 p_{k-1}(t)L_k(t)\,\mathrm{d}t = \sum_{l\in\{1:m\}}\omega_l p_{k-1}(\xi_l)L(\xi_l)$, so that

$$\|p\|_\xi^2 = \sum_{l\in\{1:m\}}\omega_l(p_{k-1}(\xi_l) + \lambda L(\xi_l))^2 = \|p_{k-1}\|_\xi^2 + \lambda^2\|L_k\|_\xi^2.$$

Combining the above two equalities proves the assertion.

Exercise 6.3 (Gauss–Radau). (i) We already know from Lemma 6.4 that $m-1 \le k_{\mathcal{Q}} \le 2m-1$. Let $p \in \mathbb{P}_{2m-2}$ and write $p = p_1(L_m - L_{m-1}) + p_2$ with $p_1 \in \mathbb{P}_{m-2}$ and $p_2 \in \mathbb{P}_{m-1}$. Owing to the L^2-orthogonality of the Legendre polynomials, the fact that the quadrature is at least of order $(m-1)$, and the definition of the Gauss–Radau nodes (which implies that $p(\xi_l) = p_2(\xi_l)$ for all $l \in \{1{:}m\}$), we infer that

$$\int_{-1}^1 p(t)\,\mathrm{d}t = \int_{-1}^1 p_2(t)\,\mathrm{d}t = \sum_{l\in\{1:m\}}\omega_l p_2(\xi_l) = \sum_{l\in\{1:m\}}\omega_l p(\xi_l).$$

Hence, $k_{\mathcal{Q}} \ge 2m-2$. The quadrature is not of higher order since it does not integrate exactly the polynomial $\frac{(L_m(t)-L_{m-1}(t))^2}{t-1}$ which is of degree $(2m-1)$ (the quadrature approximates its integral

by zero).
(ii) The polynomial $\frac{L_m(t)-L_{m-1}(t)}{t-1}L'_{m-1}(t)$ is of degree $(2m-3)$, so that it is integrated exactly by the quadrature. Since this polynomial vanishes at all the Gauss–Radau nodes except at $\xi_m = 1$, using l'Hôpital's rule we infer that

$$\int_{-1}^{1} \frac{L_m(t) - L_{m-1}(t)}{t-1} L'_{m-1}(t)\,\mathrm{d}t = \omega_m (L'_m(1) - L'_{m-1}(1)) L'_{m-1}(1) = \omega_m \frac{m^2(m-1)}{2}.$$

Moreover, since the polynomial $\big(\frac{L_m(t)-L_{m-1}(t)}{t-1}\big)'$ is of degree $(m-2)$, integrating by parts leads to

$$\int_{-1}^{1} \frac{L_m(t) - L_{m-1}(t)}{t-1} L'_{m-1}(t)\,\mathrm{d}t = \left[\frac{L_m(t) - L_{m-1}(t)}{t-1} L_{m-1}(t)\right]_{-1}^{1} = m-1.$$

Combining the above two equalities shows that $\omega_m = \frac{2}{m^2}$.
(iii) Assume $m \geq 2$ and let $l \in \{1{:}m{-}1\}$. Applying (6.3a) at $t = \xi_l$ and since $L_m(\xi_l) = L_{m-1}(\xi_l)$, we infer that $\frac{1}{m}(\xi_l^2 - 1)L'_m(\xi_l) = (\xi_l - 1)L_m(\xi_l)$. Proceeding similarly with (6.3b) leads to $L'_m(\xi_l) = mL_m(\xi_l) + \xi_l L'_{m-1}(\xi_l)$. Combining the above two equalities proves that $L'_m(\xi_l) = -L'_{m-1}(\xi_l)$.

Let us prove that $\mathcal{L}_l(t) = \frac{L_m(t)-L_{m-1}(t)}{(1-t)(t-\xi_l)} \frac{1-\xi_l}{-2L'_{m-1}(\xi_l)}$. Both functions are polynomials of degree $(m-2)$, they vanish at the $(m-2)$ interior Gauss–Lobatto nodes except at ξ_l where they take the common value 1 since $L'_m(\xi_l) - L'_{m-1}(\xi_l) = -2L'_{m-1}(\xi_l)$, as we just showed above. Hence, these two polynomials coincide identically.

Let us finally prove (6.12). Since the polynomial $\mathcal{L}_l(t)(1-t)L'_{m-1}(t)$ is of degree $(2m-3)$, it is integrated exactly by the quadrature. Using the quadrature, we infer that

$$\int_{-1}^{1} \mathcal{L}_l(t)(1-t)L'_{m-1}(t)\,\mathrm{d}t = \omega_l(1-\xi_l)L'_{m-1}(\xi_l).$$

Moreover, since the polynomial $\big(\mathcal{L}_l(t)(1-t)\big)'$ is of degree $(m-2)$, integrating by parts in time and using the L^2-orthogonality property of L_{m-1} leads to

$$\begin{aligned}
\int_{-1}^{1} \mathcal{L}_l(t)(1-t)L'_{m-1}(t)\,\mathrm{d}t &= -2\mathcal{L}_l(-1)L_{m-1}(-1) \\
&= -2\frac{L_m(-1) - L_{m-1}(-1)}{-2(1+\xi_l)} \frac{1-\xi_l}{-2L'_{m-1}(\xi_l)} L_{m-1}(-1) \\
&= \frac{1-\xi_l}{1+\xi_l} \frac{1}{L'_{m-1}(\xi_l)},
\end{aligned}$$

where we used the above expression for $\mathcal{L}_l(t)$ and that $L_m(-1) = (-1)^m$. Combining the above two equalities proves the assertion.

Exercise 6.4 (Inverse trace inequality). Using the norm equivalence from Exercise 6.2, ex-

tended to $\mathbb{R}^d$ by tensorization, we obtain

$$\begin{aligned}\|v\|_{L^2(K)}^2 &\le \sum_{\alpha\in I_{k,d}} \omega_\alpha v(a_\alpha)^2 = \sum_{i\in\{1:d\}} \sum_{\substack{\alpha\in I_{k,d}\\ \alpha_i\in\{1,k+1\}}} \omega_\alpha v(a_\alpha)^2\\ &= \frac{2}{k(k+1)} \sum_{i\in\{1:d\}} \sum_{\substack{\alpha\in I_{k,d}\\ \alpha_i\in\{1:k+1\}}} \Big(\prod_{j\neq i}\omega_{\alpha_j}\Big) v(a_\alpha)^2\\ &\le \frac{2}{k(k+1)} \sum_{i\in\{1:d\}} \left(2+\frac{1}{k}\right)^{d-1} \|v\|_{L^2(\{x_i=\pm1\})}^2\\ &= \frac{2}{k(k+1)} \left(2+\frac{1}{k}\right)^{d-1} \|v\|_{L^2(\partial K)}^2,\end{aligned}$$

with $\omega_\alpha := \prod_{j\in\{1:d\}} \omega_{\alpha_j}$.

Exercise 6.5 (Lagrange mass matrix). Let $k := n_{\rm sh}-1$. Since the set $\{L_m\}_{m\in\{0:k\}}$ is a basis of $\mathbb{P}_k$, letting $\sigma_j(p) := p(a_{j-1})$ for all $j\in\mathcal{N}$, the generalized Vandermonde matrix with entries $\mathcal{V}_{ij} := \sigma_j((\frac{2i-1}{2})^{\frac12} L_{i-1})$ is invertible. Owing to Proposition 5.5, we infer that

$$\mathcal{L}_{i-1}^{[a]}(t) = \sum_{j\in\mathcal{N}} (\mathcal{V}^{-1})_{ij} \left(\frac{2j-1}{2}\right)^{\frac12} L_{j-1}(t).$$

Hence, we have

$$\begin{aligned}\mathcal{M}_{ij} &= \int_{-1}^1 \mathcal{L}_{i-1}^{[a]}(t)\mathcal{L}_{j-1}^{[a]}(t)\,\mathrm{d}t\\ &= \sum_{m,l\in\mathcal{N}} (\mathcal{V}^{-1})_{im}(\mathcal{V}^{-1})_{jl} \left(\frac{2m-1}{2}\frac{2l-1}{2}\right)^{\frac12} \int_{-1}^1 L_{m-1}(t)L_{l-1}(t)\,\mathrm{d}t\\ &= \sum_{m,l\in\mathcal{N}} (\mathcal{V}^{-1})_{im}(\mathcal{V}^{-1})_{jl}\delta_{ml} = (\mathcal{V}^{-1}\mathcal{V}^{-\mathsf{T}})_{ij}.\end{aligned}$$

Exercise 6.6 (Canonical hybrid element). For $k=1$, the dofs define a Lagrange finite element. For $k\ge 2$, we observe that $\dim(\mathbb{P}_k) = k+1 = \operatorname{card}\Sigma$ and that a polynomial $p\in\mathbb{P}_k$ verifying $\sigma_l(p)=0$ for all $l\in\{0:k\}$ is such that $p(\pm1)=0$, so that $p(t)=(1-t^2)q(t)$ with $q\in\mathbb{P}_{k-2}$. Taking the moment of p against q yields $q=0$.
Let us verify that the shape functions are

$$\begin{aligned}\theta_0(t) &= \frac{(-1)^{k-1}}{2}(1-t)J_{k-1}^{1,0}(t),\\ \theta_l(t) &= \frac{1}{c_{l-1,1,1}}(1-t^2)J_{l-1}^{1,1}(t), \quad \forall l\in\{1:k-1\},\\ \theta_k(t) &= \frac12(1+t)J_{k-1}^{0,1}(t).\end{aligned}$$

Clearly, $\sigma_k(\theta_0)=\theta_0(1)=0$ and

$$\sigma_0(\theta_0) = \theta_0(-1) = \tfrac{(-1)^{k-1}}{2} 2J_{k-1}^{1,0}(-1) = (-1)^{k-1}(-1)^{k-1} = 1.$$

Moreover, for all $l \in \{1{:}k-1\}$, we have

$$\sigma_l(\theta_0) = \frac{(-1)^{k-1}}{2} \int_{-1}^{+1} (1-t) J_{k-1}^{1,0}(t) J_{l-1}^{1,1}(t)\,\mathrm{d}t.$$

But $J_{l-1}^{1,1} \in \mathbb{P}_{k-2} = \mathrm{span}\{J_0^{1,0}, \dots, J_{k-2}^{1,0}\}$ so that $\sigma_l(\theta_0) = 0$. In conclusion, $\sigma_l(\theta_0) = \delta_{l0}$ for all $l \in \{0{:}k\}$. A similar argument shows that $\sigma_l(\theta_k) = \delta_{lk}$ for all $l \in \{0{:}k\}$. Let $l \in \{1{:}k{-}1\}$. Then we have $\sigma_0(\theta_l) = \sigma_k(\theta_l) = 0$ by definition. Moreover, for any $l' \in \{1{:}k-1\}$, we infer that

$$\sigma_l(\theta_{l'}) = \frac{1}{c_{l-1,1,1}} \int_{-1}^{+1} (1-t)^2 J_{l'-1}^{1,1}(t) J_{l-1}^{1,1}(t)\,\mathrm{d}t = \frac{c_{l-1,1,1}}{c_{l-1,1,1}} \delta_{l'-1,l-1}.$$

Hence, $\sigma_l(\theta_{l'}) = \delta_{l',l}$. In conclusion, $\sigma_l(\theta_{l'}) = \delta_{ll'}$ for all $l' \in \{0{:}k\}$.

Exercise 6.7 ($\mathbb{Q}_{k,d}$ Lagrange). Since $\operatorname{card}\Sigma = \dim(\mathbb{Q}_{k,d}) = (k+1)^d$, we have to verify that a polynomial $q \in \mathbb{Q}_{k,d}$ vanishing at all the $\mathbb{Q}_{k,d}$ Lagrange nodes vanishes identically. We do this by induction on d. The assertion holds true for $d = 1$ (where $\mathbb{P}_{k,1} = \mathbb{Q}_{k,1}$). Let now $d \geq 2$. Using the hint, we have $q(\boldsymbol{x}) = \sum_{i_d \in \{0:k\}} q_{i_d}(x_1, \dots, x_{d-1}) x_d^{i_d}$. Consider the face $\{x_d = z_d^+\}$ of the cuboid K. This face is a cuboid in $\mathbb{R}^{d-1}$, and the $\mathbb{Q}_{k,d}$-Lagrange nodes located on this face are the $\mathbb{Q}_{k,d-1}$-Lagrange nodes of this face. Let $\boldsymbol{b}_i := (b_{i,1}, \dots, b_{i,d})^\intercal$ be one of these nodes. Let us set $\widetilde{\boldsymbol{b}}_i := (b_{i,1}, \dots, b_{i,d-1})^\intercal$. Since the function $x_d \mapsto \sum_{i_d \in \{0:k\}} q_{i_d}(\widetilde{\boldsymbol{b}}_i) x_d^{i_d}$ is in $\mathbb{Q}_{k,1}$ and vanishes at $(k+1)$ distinct nodes in $[z_d^-, z_d^+]$, it is identically zero. This shows that all the functions q_{i_d} vanish at all the Lagrange nodes of the face. By the induction hypothesis, all these functions vanish identically.

Exercise 6.8 (Bicubic Hermite). First, we have $\operatorname{card}\Sigma = \dim(\mathbb{Q}_{3,2}) = 16$. Thus, it remains to show that if $p \in \mathbb{Q}_{3,2}$ is such that all its dofs vanish, then $p = 0$. Writing p in the form $p(\boldsymbol{x}) = \sum_{i,j \in \{1:4\}} \gamma_{ij} \theta_i(x_1) \theta_j(x_2)$ where $\{\theta_1, \dots, \theta_4\}$ are the shape functions of the one-dimensional Hermite finite element, we first infer using the dofs associated with the values of p and its first-order derivatives that $\gamma_{ij} = 0$ if $i \in \{1,3\}$ or $j \in \{1,3\}$. Moreover, since $\partial^2_{x_1 x_2} p(\boldsymbol{x}) = \sum_{i,j \in \{1:4\}} \gamma_{ij} \theta_i'(x_1) \theta_j'(x_2)$, we infer using the dofs associated with the values of the second-order derivatives of p that $\gamma_{22} = \gamma_{24} = \gamma_{42} = \gamma_{44} = 0$. In conclusion, we have shown that $p = 0$.

Exercise 6.9 (Face unisolvence). Without loss of generality, we consider the face F contained in the plane $\{x_d = z_d^-\}$. Let $\{\boldsymbol{a}_i\}_{i \in \mathcal{N}_F}$ be the Lagrange nodes located on F. It is clear that if $p_{|F} = 0$, then $\sigma_i(p) := p(\boldsymbol{a}_i) = 0$ for all $i \in \mathcal{N}_F$. Let us prove the converse. Let us denote by $\widetilde{\boldsymbol{a}}_i \in \mathbb{R}^{d-1}$ the point with Cartesian components $(a_{i,1}, \dots, a_{i,d-1})$, where $(a_{i,1}, \dots, a_{i,d})$ are the Cartesian components of $\boldsymbol{a}_i$. Note that $a_{i,d} = z_d^-$ since $\boldsymbol{a}_i$ is on F. Let $p \in \mathbb{Q}_{k,d}$ and assume that $p(\boldsymbol{a}_i) = 0$ for all $i \in \mathcal{N}_F$. Since p can be written as $p(\boldsymbol{x}) = \sum_{j \in \{0:d\}} q_j(x_1, \dots, x_{d-1})(x_d - z_d^-)^j$ where $q_j \in \mathbb{Q}_{k,d-1}$, the condition $p(\boldsymbol{a}_i) = 0$, for all $i \in \mathcal{N}_F$, implies that $q_0(\widetilde{\boldsymbol{a}}_i)$ for all $i \in \mathcal{N}_F$. Consider the cuboid $\widehat{F} := \prod_{j=1}^{d-1} [z_j^-, z_j^+]$ in $\mathbb{R}^{d-1}$. Consider the dofs $\widehat{\Sigma} := \{\sigma_i(q) := q(\widetilde{\boldsymbol{a}}_i)\}_{i \in \mathcal{N}_F}$. From Proposition 6.14, we know that $(\widehat{F}, \mathbb{Q}_{k,d-1}, \widehat{\Sigma})$ is a finite element, so that we can conclude that $q_0 = 0$. This, in turn, implies that $p_{|F} = 0$.

Chapter 7

Simplicial finite elements

Exercises

Exercise 7.1 (Lagrange interpolation). Let $\mathcal{I}_K$ be the $\mathbb{P}_1$ Lagrange interpolation operator on a simplex K. Prove that $\|\mathcal{I}_K(v)\|_{C^0(K)} \leq \|v\|_{C^0(K)}$ for all $v \in C^0(K)$. (*Hint*: use the convexity of K and recall that K is closed.) Does this property hold true for $\mathbb{P}_2$ Lagrange elements?

Exercise 7.2 (Geometric identities). Prove the statements in Remark 7.6. (*Hint*: use the divergence theorem to prove (7.1).)

Exercise 7.3 (Barycentric coordinates). Let K be a simplex in $\mathbb{R}^d$. (i) Prove that $\lambda_i(\boldsymbol{x}) = 1 - \frac{|F_i|}{d|K|}\boldsymbol{n}_{K|F_i}\cdot(\boldsymbol{x} - \boldsymbol{z}_i)$ for all $\boldsymbol{x} \in K$ and all $i \in \{0{:}d\}$, and that $\nabla\lambda_i = -\frac{|F_i|}{d|K|}\boldsymbol{n}_{K|F_i}$. (ii) For all $\boldsymbol{x} \in K$, let $K_i(\boldsymbol{x})$ be the simplex obtained by joining $\boldsymbol{x}$ to the d vertices $\boldsymbol{z}_j$ with $j \neq i$. Show that $\lambda_i(\boldsymbol{x}) = \frac{|K_i(\boldsymbol{x})|}{|K|}$. (iii) Prove that $\int_K \lambda_i \,\mathrm{d}x = \frac{1}{d+1}|K|$ for all $i \in \{0{:}d\}$, and that $\int_{F_j} \lambda_i \,\mathrm{d}s = \frac{1}{d}|F_j|$ for all $j \in \{0{:}d\}$ with $j \neq i$, and $\int_{F_i} \lambda_i \,\mathrm{d}s = 0$. (*Hint*: consider an affine mapping from K to the unit simplex.) (iv) Prove that if $\boldsymbol{h} \in \mathbb{R}^d$ satisfies $D\lambda_i(\boldsymbol{h}) = 0$ for all $i \in \{1{:}d\}$, then $\boldsymbol{h} = \boldsymbol{0}$.

Exercise 7.4 (Space $\mathbb{P}_{k,d}$). (i) Give a basis for $\mathbb{P}_{2,d}$ for $d \in \{1, 2, 3\}$. (ii) Show that any polynomial $p \in \mathbb{P}_{k,d}$ can be written in the form $p(x_1, \ldots, x_d) = r(x_1, \ldots, x_{d-1}) + x_d q(x_1, \ldots, x_d)$, with unique polynomials $r \in \mathbb{P}_{k,d-1}$ and $q \in \mathbb{P}_{k-1,d}$. (iii) Determine the dimension of $\mathbb{P}_{k,d}$. (*Hint*: by induction on d.) (iv) Let K be a simplex in $\mathbb{R}^d$. Let F_0 be the face of K opposite to the vertex $\boldsymbol{z}_0$. Prove that if $p \in \mathbb{P}_{k,d}$ satisfies $p_{|F_0} = 0$, then there is $q \in \mathbb{P}_{k-1,d}$ s.t. $p = \lambda_0 q$. (*Hint*: write the Taylor expansion of p at $\boldsymbol{z}_d$ and use (7.2) with $\boldsymbol{z}_d$ playing the role of $\boldsymbol{z}_0$.) (v) Prove that $\{\lambda_0^{\beta_0} \ldots \lambda_d^{\beta_d} \mid \beta_0 + \ldots + \beta_d = k\}$ is a basis of $\mathbb{P}_{k,d}$.

Exercise 7.5 (Nodes of simplicial Lagrange FE). Let K be a simplex in $\mathbb{R}^d$, and consider the set of nodes $\{\boldsymbol{a}_i\}_{i\in\mathcal{N}}$ with barycentric coordinates $\left(\frac{i_0}{k}, \ldots, \frac{i_d}{k}\right)$, $\forall i_0, \ldots, i_d \in \{0{:}k\}$ with $i_0 + \ldots + i_d = k$. (i) Prove that the number of nodes located on any one-dimensional edge of K is $(k+1)$ in any dimension $d \geq 2$. (ii) Prove that the number of nodes located on any $(d-1)$-dimensional face of K is the dimension of $\mathbb{P}_{k,d-1}$. (iii) Prove that if $k \leq d$, all the nodes are located on the boundary of K.

Exercise 7.6 (Hierarchical basis). Let $k \geq 1$ and let $\{\theta_0, \ldots, \theta_k\}$ be a hierarchical basis of $\mathbb{P}_{k,1}$. Let $\{\lambda_0, \ldots, \lambda_d\}$ be a basis of $\mathbb{P}_{1,d}$ and assume that $\lambda_i : \mathbb{R}^d \to \mathbb{R}$ is surjective for all $i \in \{0{:}d\}$ (i.e., λ_i is not constant). (i) Show that the functions (mapping $\mathbb{R}^d$ to $\mathbb{R}$) $\{\theta_0(\lambda_i), \ldots, \theta_k(\lambda_i)\}$ are linearly

independent for all $i \in \{0{:}d\}$. (*Hint*: consider a linear combination $\sum_{l\in\{0:k\}} \alpha_l\theta_l(\lambda_i) \in \mathbb{P}_{k,d}$ and prove that the polynomial $\sum_{l\in\{0:k\}} \alpha_l\theta_l \in \mathbb{P}_{k,1}$ vanishes at $(k+1)$ distinct points.) (ii) Show that the functions (mapping $\mathbb{R}^d$ to $\mathbb{R}$) from the set $S_{k,d} \coloneqq \{\theta_{\alpha_1}(\lambda_1)\dots\theta_{\alpha_d}(\lambda_d) \mid (\alpha_1,\dots\alpha_d) \in \mathbb{N}^d,\ |\alpha| \le k\}$ are linearly independent. (*Hint*: by induction on d.) (iii) Show that $(S_{k,d})_{k\ge0}$ is a hierarchical polynomial basis, i.e., $S_{k,d} \subset S_{k+1,d}$ and $S_{k,d}$ is basis of $\mathbb{P}_{k,d}$. (*Note*: the $(d+1)$ vertices of K do not play here the same role.)

Exercise 7.7 (Cubic Hermite triangle). Let K be a triangle with vertices $\{\boldsymbol{z}_0,\boldsymbol{z}_1,\boldsymbol{z}_2\}$. Set $\Sigma \coloneqq \{p(\boldsymbol{z}_i), \partial_{x_1}p(\boldsymbol{z}_i), \partial_{x_2}p(\boldsymbol{z}_i)\}_{0\le i\le2} \cup \{p(\boldsymbol{a}_K)\}$, where $\boldsymbol{a}_K$ is a point inside K. Show that $(K,\mathbb{P}_{3,2},\Sigma)$ is a finite element. (*Hint*: show that any $p \in \mathbb{P}_{3,2}$ for which all the dofs vanish is identically zero on the three edges of K and infer that $p = c\lambda_0\lambda_1\lambda_2$ for some $c \in \mathbb{R}$.)

Exercise 7.8 ($\mathbb{P}_{2,d}$ canonical hybrid FE). Compute the shape functions of the $\mathbb{P}_{2,d}$ canonical hybrid finite element for the unit simplex for $d = 1$ and $d = 2$ (provide an expression using the Cartesian coordinates and another one using the barycentric coordinates).

Exercise 7.9 ($\mathbb{P}_{4,2}$ Lagrange). Using the Lagrange nodes defined as in Proposition 7.11, give the expression of the $\mathbb{P}_{4,2}$ Lagrange shape functions in terms of the barycentric coordinates.

Exercise 7.10 (Quadratic Crouzeix–Raviart). Let K be the unit simplex. Let $\alpha \in (0,1)$. Let $\boldsymbol{g}_1 \coloneqq (\alpha,0)$, $\boldsymbol{g}_2 \coloneqq (1-\alpha,0)$, $\boldsymbol{g}_3 \coloneqq (1-\alpha,\alpha)$, $\boldsymbol{g}_4 \coloneqq (\alpha,1-\alpha)$, $\boldsymbol{g}_5 \coloneqq (0,1-\alpha)$, $\boldsymbol{g}_6 \coloneqq (0,\alpha)$. (i) Compute $\lambda_0(\boldsymbol{g}_j)^2 + \lambda_1(\boldsymbol{g}_j)^2 + \lambda_2(\boldsymbol{g}_j)^2$ for all $j \in \{1{:}6\}$, where $\lambda_0,\lambda_1,\lambda_2$ are the barycentric coordinates of K. (ii) Let $\sigma_j \in \mathcal{L}(\mathbb{P}_{2,2};\mathbb{R})$ be defined by $\sigma_j(p) \coloneqq p(\boldsymbol{g}_j)$ for all $p \in \mathbb{P}_{2,2}$ and $j \in \{1{:}6\}$. Let $\Sigma \coloneqq \{\sigma_j\}_{j\in\{1:6\}}$. Is the triple $(K,\mathbb{P}_{2,2},\Sigma)$ a finite element? (iii) Let F_i, $i \in \{0{:}2\}$, be one of the three faces of K. Let $\boldsymbol{T}_{F_i} : [-1,1] \to F_i$ be one of the two affine mappings that realize a bijection between $[-1,1]$ and F_i. Let $\{q_0,q_1\}$ be a basis of $\mathbb{P}_{1,1}$. Let $\varpi_{2i+k} \in \mathcal{L}(\mathbb{P}_{2,2};\mathbb{R})$, $i \in \{0{:}2\}$, $k \in \{0{:}1\}$, be defined by $\varpi_{2i+k}(p) \coloneqq \frac{1}{|F_i|}\int_{F_i}(q_k\circ\boldsymbol{T}_{F_i}^{-1})p\,\mathrm{d}s$ for all $p \in \mathbb{P}_{2,2}$. Let $\Sigma \coloneqq \{\varpi_j\}_{j\in\{0:5\}}$. Is the triple $(K,\mathbb{P}_{2,2},\Sigma)$ a finite element? (*Hint*: consider the points $\boldsymbol{T}_{F_i}(\xi_k)$, $i \in \{0{:}2\}$, $k \in \{0{:}1\}$, where ξ_0, ξ_1 are the two nodes of the Gauss–Legendre quadrature of order 3, then use Step (ii).)

Solution to exercises

Exercise 7.1 (Lagrange interpolation). Since K is convex and closed and since $\mathcal{I}_K(v)$ is affine, $\mathcal{I}_K(v)$ reaches its extrema at a vertex of K, where its value coincides with that of v. This property fails for piecewise quadratic functions since the basis functions can take negative values.

Exercise 7.2 (Geometric identities). It suffices to prove that the family $\{\boldsymbol{n}_{K|F_i}\}_{i\in\{1:d\}}$ is linearly independent. Let $\boldsymbol{h} \in \mathbb{R}^d$ be s.t. $\sum_{i\in\{1:d\}} h_i\boldsymbol{n}_{K|F_i} = \boldsymbol{0}$. Taking the $\ell^2(\mathbb{R}^d)$-inner product with $(\boldsymbol{z}_j - \boldsymbol{z}_0)$ and observing that $\boldsymbol{n}_{K|F_i}\cdot(\boldsymbol{z}_j-\boldsymbol{z}_0) = \delta_{ij}\frac{d|K|}{|F_i|}$, we infer that $h_i = 0$ for all $i \in \{1{:}d\}$. Let us now prove (7.1) Let $\boldsymbol{h} \in \mathbb{R}^d$. We observe that

$$
\begin{aligned}
0 = \int_K \nabla\cdot\boldsymbol{h}\,\mathrm{d}x &= \int_{\partial K} \boldsymbol{h}\cdot\boldsymbol{n}_K\,\mathrm{d}s \\
&= \sum_{i\in\{0:d\}} |F_i|\boldsymbol{h}\cdot\boldsymbol{n}_{K|F_i} = \boldsymbol{h}\cdot\left(\sum_{i\in\{0:d\}} |F_i|\boldsymbol{n}_{K|F_i}\right),
\end{aligned}
$$

yielding the first geometric identity. For the second one, integration by parts yields for all $\boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathbb{R}^d$,

$$\begin{aligned}|K|\boldsymbol{h}_1\cdot\boldsymbol{h}_2 &= \int_K \nabla(\boldsymbol{h}_1\cdot\boldsymbol{x})\cdot\boldsymbol{h}_2\,\mathrm{d}x = \sum_{i\in\{0:\,d\}}\int_{F_i}(\boldsymbol{h}_1\cdot\boldsymbol{x})(\boldsymbol{h}_2\cdot\boldsymbol{n}_{K|F_i})\,\mathrm{d}s\\ &= \sum_{i\in\{0:\,d\}}|F_i|(\boldsymbol{h}_1\cdot\boldsymbol{c}_{F_i})(\boldsymbol{h}_2\cdot\boldsymbol{n}_{K|F_i})\\ &= \boldsymbol{h}_1\cdot\left(\sum_{i\in\{0:\,d\}}|F_i|(\boldsymbol{c}_{F_i}-\boldsymbol{c}_K)\otimes\boldsymbol{n}_{K|F_i}\right)\cdot\boldsymbol{h}_2,\end{aligned}$$

where we used that $\nabla(\boldsymbol{h}_1\cdot\boldsymbol{x}) = \boldsymbol{h}_1$, the definition of $\boldsymbol{c}_{F_i}$, and the first geometric identity to introduce $\boldsymbol{c}_K$. Since the vectors $\boldsymbol{h}_1, \boldsymbol{h}_2$ are arbitrary in $\mathbb{R}^d$, we infer the second geometric identity.

Exercise 7.3 (Barycentric coordinates). (i) Since λ_i is affine and constant on F_i, its gradient is constant and collinear to $\boldsymbol{n}_{K|F_i}$. Using $\lambda_i(\boldsymbol{z}_i) = 1$, we infer that $\lambda_i(\boldsymbol{x}) = 1 - c_i\boldsymbol{n}_{K|F_i}\cdot(\boldsymbol{x}-\boldsymbol{z}_i)$. We obtain c_i by using that $|K| = \frac{1}{d}|F_i|(\boldsymbol{n}_{K|F_i}\cdot(\boldsymbol{z}_j-\boldsymbol{z}_i))$ for all $j\neq i$. The expression for the gradient follows immediately.
(ii) The function $\frac{|K_i(\boldsymbol{x})|}{|K|}$ is in $\mathbb{P}_{1,d}$ and coincides with $\lambda_i(\boldsymbol{x})$ at the $(d+1)$ vertices of K, so both functions coincide everywhere. Another way to look at this problem consists of observing that $\sum_{i\in\{1:\,d\}}(\boldsymbol{z}_i-\boldsymbol{z}_0)\lambda_i(\boldsymbol{x}) = \boldsymbol{x}-\boldsymbol{z}_0$, i.e.,

$$\begin{pmatrix} z_{1,1}-z_{0,1} & \cdots & z_{d,1}-z_{0,1}\\ \vdots & \ddots & \vdots\\ z_{1,d}-z_{0,d} & \cdots & z_{d,d}-z_{0,d}\end{pmatrix}\begin{pmatrix}\lambda_1(\boldsymbol{x})\\ \vdots\\ \lambda_d(\boldsymbol{x})\end{pmatrix} = \begin{pmatrix} x_1-z_{0,1}\\ \vdots\\ x_d-z_{0,d}\end{pmatrix},$$

where $(z_{i,1},\dots,z_{i,d})$ are the Cartesian coordinates of $\boldsymbol{z}_i$. Cramer's rule implies that

$$\lambda_i(\boldsymbol{x}) = \frac{\det(\boldsymbol{z}_1-\boldsymbol{z}_0,\dots,\boldsymbol{x}-\boldsymbol{z}_0,\dots,\boldsymbol{z}_d-\boldsymbol{z}_0)}{\det(\boldsymbol{z}_1-\boldsymbol{z}_0,\dots,\boldsymbol{z}_d-\boldsymbol{z}_0)},$$

where $(\boldsymbol{z}_1-\boldsymbol{z}_0,\dots,\boldsymbol{x}-\boldsymbol{z}_0,\dots,\boldsymbol{z}_d-\boldsymbol{z}_0)$ is the matrix with the column vectors $\boldsymbol{z}_1-\boldsymbol{z}_0$, …, $\boldsymbol{x}-\boldsymbol{z}_0$, …, $\boldsymbol{z}_d-\boldsymbol{z}_0$ with the vector $\boldsymbol{x}-\boldsymbol{z}_0$ in the i-th column. Since $|K| = |\det(\boldsymbol{z}_1-\boldsymbol{z}_0,\dots,\boldsymbol{z}_d-\boldsymbol{z}_0)|$ and $|K_i(\boldsymbol{x})| = |\det(\boldsymbol{z}_1-\boldsymbol{z}_0,\dots,\boldsymbol{x}-\boldsymbol{z}_0,\dots,\boldsymbol{z}_d-\boldsymbol{z}_0)|$, we infer that $\lambda_i(\boldsymbol{x}) = \frac{|K_i(\boldsymbol{x})|}{|K|}$.
(iii) Consider an affine transformation, say $\boldsymbol{T}_K$, mapping K to the unit simplex, say $\widehat{K}$. Then $\int_K\lambda_i\,\mathrm{d}x = \frac{|K|}{|\widehat{K}|}\int_{\widehat{K}}\widehat{\lambda}_i\,\mathrm{d}\widehat{x}$ and $\widehat{\lambda}_i$ is the barycentric coordinate associated with the vertex $\boldsymbol{T}_K(\boldsymbol{z}_i)$ in $\widehat{K}$. A direct computation shows that $|\widehat{K}| = \frac{1}{d!}$ and $\int_{\widehat{K}}\widehat{\lambda}_i\,\mathrm{d}\widehat{x} = \frac{1}{(d+1)!}$. Hence, $\int_K\lambda_i\,\mathrm{d}x = \frac{1}{d+1}|K|$. The proof for the integral of λ_i on the faces of K is similar.
(iv) Writing $\boldsymbol{h} := \sum_{j\in\{1:\,d\}}h_j(\boldsymbol{z}_j-\boldsymbol{z}_0)$, we infer that for all $i,j\in\{1{:}d\}$,

$$\begin{aligned}D\lambda_i(\boldsymbol{h}) &= \sum_{j\in\{1:\,d\}}h_jD\lambda_i(\boldsymbol{z}_j-\boldsymbol{z}_0)\\ &= \sum_{j\in\{1:\,d\}}h_j(\lambda_i(\boldsymbol{z}_j)-\lambda_i(\boldsymbol{z}_0)) = \sum_{j\in\{1:\,d\}}h_j\delta_{ij} = h_i,\end{aligned}$$

since λ_i is affine, $\lambda_i(\boldsymbol{z}_j) = \delta_{ij}$ and $\lambda_i(\boldsymbol{z}_0) = 0$. Hence, $\boldsymbol{h} = \boldsymbol{0}$.

Exercise 7.4 (Space $\mathbb{P}_{k,d}$). (i) For $d=1$, a basis is

$$\{1, x, x^2\}.$$

For $d=2$, a basis is

$$\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2\}.$$

For $d=3$, a basis is

$$\{1, x_1, x_2, x_3, x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, x_3^2\}.$$

(ii) Writing $\boldsymbol{\alpha} := (\boldsymbol{\alpha}', \alpha_d)$ with $\boldsymbol{\alpha}' \in \mathbb{N}^{d-1}$ for a multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$ and writing $\boldsymbol{x} := (\boldsymbol{x}', x_d)$ with $\boldsymbol{x}' \in \mathbb{R}^{d-1}$ for $\boldsymbol{x} \in \mathbb{R}^d$, we have

$$\begin{aligned} p(\boldsymbol{x}) &= \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{k,d},\alpha_d=0} a_{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}} + \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{k,d},\alpha_d\ge 1} a_{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}} \\ &= \sum_{\boldsymbol{\alpha}'\in\mathcal{A}_{k,d-1}} a_{(\boldsymbol{\alpha}',0)}(\boldsymbol{x}')^{\boldsymbol{\alpha}'} + x_d\left(\sum_{\boldsymbol{\beta}\in\mathcal{A}_{k-1,d}} a_{\boldsymbol{\beta}+\boldsymbol{\gamma}_d}\boldsymbol{x}^{\boldsymbol{\beta}}\right), \end{aligned}$$

where in the second sum we introduced the multi-index $\boldsymbol{\gamma}_d := (0,\dots,0,1) \in \mathbb{N}^d$. This means that $p(\boldsymbol{x}) = r(\boldsymbol{x}') + x_d q(\boldsymbol{x})$ with $r \in \mathbb{P}_{k,d-1}$ and $q \in \mathbb{P}_{k-1,d}$. Uniqueness of r and q results from the fact that if $r + x_d q$ vanishes identically, then taking $x_d = 0$ first yields $r = 0$ so that $x_d q = 0$, whence we infer that $q = 0$.
(iii) For $d=1$, $\dim \mathbb{P}_{k,1} = (k+1) = \binom{k+1}{1}$. Assume that for $d \ge 2$, $\dim \mathbb{P}_{k,d-1} = \binom{k+d-1}{d-1}$. For $k=1$, $\dim \mathbb{P}_{1,d} = (d+1) = \binom{1+d}{d}$ since there is at most one nonzero index α_j for $j \in \{1{:}d\}$. Owing to Step (ii), we have $\dim \mathbb{P}_{k,d} = \dim \mathbb{P}_{k,d-1} + \dim \mathbb{P}_{k-1,d}$, so that

$$\dim \mathbb{P}_{k,d} = \binom{k+d-1}{d-1} + \binom{k+d-1}{d} = \binom{k+d}{d}.$$

(iv) Writing (7.2) with $\boldsymbol{z}_d$ playing the role of $\boldsymbol{z}_0$, we infer that $\boldsymbol{h} := \boldsymbol{x} - \boldsymbol{z}_d = \sum_{i\in\{0:d-1\}} \lambda_i(\boldsymbol{x})(\boldsymbol{z}_i - \boldsymbol{z}_d)$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Writing the Taylor expansion of p at $\boldsymbol{z}_d$ of order k, we infer that

$$p(\boldsymbol{x}) = p(\boldsymbol{z}_d) + \sum_{l\in\{1:k\}} \frac{1}{l!} D^l p(\boldsymbol{z}_d)(\boldsymbol{h},\dots,\boldsymbol{h}) = \sum_{l\in\{1:k\}} \frac{1}{l!} D^l p(\boldsymbol{z}_d)(\boldsymbol{h},\dots,\boldsymbol{h}),$$

since $p(\boldsymbol{z}_d) = 0$. Let $\mathcal{M}_l := \{0{:}d-1\}^l$. Using the multilinearity of the Fréchet derivative, we infer that

$$D^l p(\boldsymbol{z}_d)(\boldsymbol{h},\dots,\boldsymbol{h}) = \sum_{\mu\in\mathcal{M}_l} \lambda_{\mu_1}\dots\lambda_{\mu_l}\Omega_{l,\mu},$$

where $\Omega_{l,\mu} = D^l p(\boldsymbol{z}_d)(\boldsymbol{z}_{\mu_1} - \boldsymbol{z}_d, \dots, \boldsymbol{z}_{\mu_l} - \boldsymbol{z}_d)$ is a real number. Since $p_{|F_0} \equiv 0$, we infer that $D^l p(\boldsymbol{z}_d)(\boldsymbol{z}_{\mu_1} - \boldsymbol{z}_d, \dots, \boldsymbol{z}_{\mu_l} - \boldsymbol{z}_d) = 0$ if all the indices $\mu_1,\dots,\mu_l$ are not zero, because $D^l p(\boldsymbol{z}_d)(\boldsymbol{z}_{\mu_1} - \boldsymbol{z}_d, \dots, \boldsymbol{z}_{\mu_l} - \boldsymbol{z}_d)$ is a tangential derivative along F_0 in this case. Let $\mathcal{M}_l^* := \{\mu \in \mathcal{M}_l \mid \exists j_\mu \in \{1{:}l\},\ \mu_{j_\mu} = 0\}$. The above argument implies that

$$D^l p(\boldsymbol{z}_d)(\boldsymbol{h},\dots,\boldsymbol{h}) = \sum_{\mu\in\mathcal{M}_l^*} \lambda_{\mu_1}\dots\lambda_{\mu_l}\Omega_{l,\mu} = \lambda_0 \sum_{\mu\in\mathcal{M}_l^*} \left(\prod_{j\ne j_\mu} \lambda_{\mu_j}\right)\Omega_{l,\mu}.$$

Hence, $p = \lambda_0 q$ with $q = \sum_{l\in\{1:k\}} \frac{1}{l!} \sum_{\mu\in\mathcal{M}_l^*} \left(\prod_{j\neq j_\mu} \lambda_{\mu_j}\right) \Omega_{l,\mu}$, and since the barycentric coordinates are affine functions, we infer that $q \in \mathbb{P}_{k-1,d}$.
(v) It suffices to prove linear independence since

$$\mathrm{card}(\{(\beta_0,\dots,\beta_d) \in \mathbb{N}^{d+1} \mid \beta_0 + \dots + \beta_d = k\}) = \binom{k+d}{k} = \dim(\mathbb{P}_{k,d}).$$

Assume that $\sum_\beta \mu_{\beta_0\dots\beta_d} \lambda_0^{\beta_0} \dots \lambda_d^{\beta_d} = 0$. Restricting this function to the face F_0 where λ_0 vanishes identically and using induction on d, we infer that $\mu_{\beta_0\dots\beta_d} = 0$ whenever $\beta_0 = 0$. If $k = 1$, we infer that $\mu_{10\dots0}\lambda_0$ vanishes identically, so that $\mu_{10\dots0} = 0$. If $k \geq 2$, we can factor out λ_0 and use induction on k to conclude that $\mu_{\beta_0\dots\beta_d} = 0$ for all $\beta_0 \geq 1$.

Exercise 7.5 (Nodes of simplicial Lagrange FE). (i) Consider a one-dimensional edge of K. There are two distinct integers $j_1, j_2 \in \{0{:}d\}$ such that this edge connects the vertices $\boldsymbol{z}_{j_1}$ and $\boldsymbol{z}_{j_2}$. Then the nodes located on the edge correspond to setting $i_j := 0$ for all $j \in \{0{:}d\} \setminus \{j_1, j_2\}$. In other words, such nodes correspond to the choices $j_1, j_2 \in \{0{:}d\}$ and $j_1 + j_2 = k$, and there are $(k+1)$ such choices.
(ii) Consider now a $(d-1)$-dimensional face of K. There is an integer $j \in \{0{:}d\}$ such that all the nodes located on this face are such that $i_j = 0$, and there are $\binom{k+d-1}{d-1}$ choices for these nodes, which is the dimension of $\mathbb{P}_{k,d-1}$.
(iii) Assume $k \leq d$. This means that at least one index i_j, for $j \in \{0{:}d\}$, vanishes. Then the corresponding node is located on the $(d-1)$-dimensional face of K opposite to the vertex $\boldsymbol{z}_j$.

Exercise 7.6 (Hierarchical basis). (i) Let $\alpha_0,\dots,\alpha_k \in \mathbb{R}$ and assume that $\alpha_0\theta_0(\lambda_i(\boldsymbol{x})) + \dots + \alpha_k\theta_k(\lambda_i(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Let $x_0,\dots x_k$ be $(k+1)$ distinct real numbers. Since $\lambda_i : \mathbb{R}^d \to \mathbb{R}$ is surjective, there are $(k+1)$ points $\boldsymbol{x}_0,\dots \boldsymbol{x}_k$ in $\mathbb{R}^d$ such that $\lambda_i(\boldsymbol{x}_l) = x_l$ for all $l \in \{0{:}k\}$. Hence, $\alpha_0\theta_0(x_l) + \dots + \alpha_k\theta_k(x_l) = 0$ for all $l \in \{0{:}k\}$. This means that the polynomial $\alpha_0\theta_0 + \dots + \alpha_k\theta_k \in \mathbb{P}_{k,1}$ vanishes at $(k+1)$ distinct points. Hence, this polynomial vanishes identically. Since $\{\theta_0,\dots,\theta_k\}$ is a basis of $\mathbb{P}_{k,1}$, we infer that $\alpha_0 = \dots = \alpha_k = 0$.
(ii) We prove the statement by induction over $d \geq 1$. The statement has been proved for $d = 1$ in Step (i). Assume now that $d \geq 2$. Recall the set $\mathcal{A}_{k,d}$ from §7.3. Let $\{a_\alpha\}_{\alpha\in\mathcal{A}_{k,d}}$ and assume that

$$\sum_{\alpha\in\mathcal{A}_{k,d}} a_\alpha \theta_{\alpha_1}(\lambda_1(\boldsymbol{x}))\dots\theta_{\alpha_d}(\lambda_d(\boldsymbol{x})) = 0, \qquad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

We infer that

$$0 = \sum_{\alpha_d\in\{0:k\}} \theta_{\alpha_d}(\lambda_d(\boldsymbol{x})) \times \left(\sum_{|\alpha|_{d-1}\leq k-\alpha_d} a_\alpha \theta_{\alpha_1}(\lambda_1(\boldsymbol{x}))\dots\theta_{\alpha_{d-1}}(\lambda_{d-1}(\boldsymbol{x})) \right),$$

with $|\alpha|_{d-1} := \alpha_1 + \dots + \alpha_{d-1}$. Since $\{\lambda_0\dots\lambda_d\}$ is a basis of $\mathbb{P}_{1,d}$, there are $\boldsymbol{z}_0,\dots,\boldsymbol{z}_d \in \mathbb{R}^d$ such that

$$\boldsymbol{x} = \sum_{i\in\{0:d\}} \boldsymbol{z}_i \lambda_i(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{R}^d,$$

and this representation is unique. Let $y_1,\dots,y_{d-1},x$ be d arbitrary real numbers and let $\boldsymbol{x} := \boldsymbol{z}_d x + \sum_{i\in\{1:d-1\}} \boldsymbol{z}_i y_i$. Then $\lambda_1(\boldsymbol{x}) = y_1,\dots,\lambda_{d-1}(\boldsymbol{x}) = y_{d-1}, \lambda_d(\boldsymbol{x}) = x$. We infer that

$$0 = \sum_{\alpha_d\in\{0:k\}} \theta_{\alpha_d}(x) \times \left(\sum_{|\alpha|_{d-1}\leq k-\alpha_d} a_\alpha \theta_{\alpha_1}(y_1)\dots\theta_{\alpha_{d-1}}(y_{d-1}) \right),$$

for all $x \in \mathbb{R}$. Since $\{\theta_0, \dots, \theta_k\}$ is a basis of $\mathbb{P}_{k,1}$, we conclude that

$$0 = \sum_{\alpha_1+\dots+\alpha_{d-1} \le k-\alpha_d} a_\alpha \theta_{\alpha_1}(y_1) \dots \theta_{\alpha_{d-1}}(y_{d-1}),$$

for all $(y_1, \dots y_{d-1}) \in \mathbb{R}^{d-1}$ and all $\alpha_d \in \{0{:}d\}$. Let H be the affine hyperplane passing through the points $\boldsymbol{z}_0, \dots, \boldsymbol{z}_{d-1}$ and let $\boldsymbol{T} : \mathbb{R}^{d-1} \to H$ be such that

$$\boldsymbol{T}(\boldsymbol{y}) := \boldsymbol{z}_0 + \sum_{i\in\{1:d-1\}} y_i(\boldsymbol{z}_i - \boldsymbol{z}_0) = \Big(1 - \sum_{i\in\{1:d-1\}} y_i\Big)\boldsymbol{z}_0 + \sum_{i\in\{1:d-1\}} y_i \boldsymbol{z}_i.$$

The identity $\boldsymbol{x} = \sum_{i\in\{0:d\}} \boldsymbol{z}_i \lambda_i(\boldsymbol{x})$ obtained above implies that $\lambda_0(\boldsymbol{T}(\boldsymbol{y})) = 1 - \sum_{i\in\{1:d-1\}} y_i$ and $\lambda_i(\boldsymbol{T}(\boldsymbol{y})) = y_i$ for all $i \in \{1{:}d{-}1\}$ (note in passing that $\lambda_d(\boldsymbol{T}(\boldsymbol{y})) = 0$). Let us set $\tilde\lambda_i(\boldsymbol{y}) := \lambda_i(\boldsymbol{T}(\boldsymbol{y}))$. It is clear that the set $\{\tilde\lambda_0, \dots, \tilde\lambda_{d-1}\}$ is linearly independent. Moreover, this set has cardinality $d = \dim \mathbb{P}_{1,d-1}$, so that $\{\tilde\lambda_0, \dots, \tilde\lambda_{d-1}\}$ is a basis of $\mathbb{P}_{1,d-1}$. Finally, the above argument shows that

$$0 = \sum_{|\alpha|_{d-1} \le k-\alpha_d} a_\alpha \theta_{\alpha_1}(\tilde\lambda_1(\boldsymbol{y})) \dots \theta_{\alpha_{d-1}}(\tilde\lambda_{d-1}(\boldsymbol{y})),$$

for all $(y_1, \dots y_{d-1}) \in \mathbb{R}^{d-1}$ and all $\alpha_d \in \{0{:}d\}$. The induction hypothesis implies that $a_\alpha = 0$.
(iii) We have proved in Step (ii) that $S_{k,d}$ is linearly independent. Moreover, $\operatorname{card}(S_{k,d}) = \operatorname{card}(\mathcal{A}_{k,d}) = \binom{k+d}{d} = \dim(\mathbb{P}_{k,d})$. Hence, $S_{k,d}$ is a basis of $\mathbb{P}_{k,d}$. Finally, it is clear that $S_{k,d} \subset S_{k+1,d}$, i.e., $S_{k,d}$ is a hierarchical basis of $\mathbb{P}_{k,d}$.

Exercise 7.7 (Cubic Hermite triangle). We first observe that $\operatorname{card}\Sigma = \dim \mathbb{P}_{3,2} = 10$. Let $p \in \mathbb{P}_{3,2}$ be such that all its dofs vanish. Restricting p to the face F_i of K, $i \in \{0,1,2\}$, we obtain a polynomial in $\mathbb{P}_{3,1}$ that vanishes at the two endpoints as well as its derivative. Hence, p vanishes identically on F_i. Using the result for F_0 implies that $p = \lambda_0 q_0$ with $q_0 \in \mathbb{P}_{2,2}$. Since q_0 vanishes identically on F_1, $q_0 = \lambda_1 q_1$ with $q_1 \in \mathbb{P}_{1,2}$, and reasoning similarly for F_2 yields $p = c\lambda_0\lambda_1\lambda_2$ for some $c \in \mathbb{R}$. Evaluating p at the interior point $\boldsymbol{a}_K$ for which all the barycentric coordinates are nonzero, we infer that $c = 0$.

Exercise 7.8 ($\mathbb{P}_{2,d}$ canonical hybrid FE). (i) Let us start with the unit simplex in dimension one, i.e., $K := [0,1]$. Let us set $\boldsymbol{z}_0 := 0$ and $\boldsymbol{z}_1 := 1$. Let $\theta_0(x) := (1-x)(1+ax)$. Observe that $\theta_0(\boldsymbol{z}_0) = 1$ and $\theta_0(\boldsymbol{z}_1) = 0$. We now compute a so that $\int_0^1 \theta_0(x)\,\mathrm{d}x = 0$ (here we take $\mathbb{P}_{0,1} := \operatorname{span}\{1\}$). Then $1 - \frac12 + a(\frac12 - \frac13) = 0$ gives $a = -3$. Hence, $\theta_0(x) = (1-x)(1-3x)$. Similarly, by symmetry, we have $\theta_1(x) = x(3x-2)$ (just replace x by $1-x$ in the expression of $\theta_0(x)$). For the third shape function, we have $\theta_2(x) = ax(1-x)$. The constraint $a\int_0^1 x(1-x)\,\mathrm{d}x = 1$ gives $a(\frac12 - \frac13) = 1$. Hence, $a = 6$ and $\theta_2(x) = 6x(1-x)$. In terms of the barycentric coordinates $\lambda_0(x) := 1-x$ and $\lambda_1(x) := x$, we obtain

$$\theta_0 = \lambda_0(3\lambda_0 - 2), \quad \theta_1 = \lambda_1(3\lambda_1 - 2), \quad \theta_2 = 6\lambda_0\lambda_1.$$

(ii) Here, we take again $\mathbb{P}_{0,2} := \operatorname{span}\{1\}$. Let us set $\boldsymbol{z}_0 := (0,0)$, $\boldsymbol{z}_1 := (1,0)$, and $\boldsymbol{z}_2 := (0,1)$. Setting $\theta_0(\boldsymbol{x}) := (1-x_1-x_2)(1+ax_2+bx_2)$, we observe that $\int_{F_0} \theta_0(\boldsymbol{x})\,\mathrm{d}x = 0$, $\theta_0(\boldsymbol{z}_1) = 0$, $\theta_0(\boldsymbol{z}_2) = 0$, and $\theta_0(\boldsymbol{z}_0) = 1$. We must also have $\int_{F_2} \theta_0(\boldsymbol{x})\,\mathrm{d}x = 0 = \int_0^1 (1-x_1)(1+ax_1)\,\mathrm{d}x_1$. Hence, $a = -3$. Similarly, we infer that $b = -3$, which proves that $\theta_0(\boldsymbol{x}) = (1-x_1-x_2)(1-3x_1-3x_2)$. By symmetry, we obtain $\theta_1(\boldsymbol{x}) = x_1(3x_1-2)$ (replace $1-x_1-x_2$ by x_1, or just do the computation) and $\theta_2(\boldsymbol{x}) = x_2(3x_2-2)$ (replace $1-x_1-x_2$ by x_2, or just do the computation). We now

compute $\theta_3(\boldsymbol{x}) = ax_1x_2$ with the constraint $\frac{1}{|F_0|}\int_{F_0}\theta_3(\boldsymbol{x})\,\mathrm{d}x = 1$. Hence, $a\int_0^1 x_1(1-x_1)\,\mathrm{d}x_1 = 1$, which gives $a = 6$, i.e., $\theta_3(\boldsymbol{x}) = 6x_1x_2$. By symmetry, we obtain $\theta_4(\boldsymbol{x}) = 6x_2(1-x_1-x_2)$ and $\theta_5(\boldsymbol{x}) = 6x_1(1-x_1-x_2)$. In terms of the barycentric coordinates $\lambda_0(\boldsymbol{x}) := 1-x_1-x_2$, $\lambda_1(\boldsymbol{x}) := x_1$, and $\lambda_2(\boldsymbol{x}) := x_2$, we obtain

$$\begin{aligned}&\theta_0 = \lambda_0(3\lambda_0-2), &&\theta_1 = \lambda_1(3\lambda_1-2), &&\theta_2 = \lambda_2(3\lambda_2-2),\\ &\theta_3 = 6\lambda_1\lambda_2, &&\theta_4 = 6\lambda_0\lambda_2, &&\theta_5 = 6\lambda_0\lambda_1.\end{aligned}$$

Exercise 7.9 ($\mathbb{P}_{4,2}$ Lagrange). Let us use the notation from Proposition 7.11, that is, the Lagrange nodes are defined by $\boldsymbol{a}_\alpha := \boldsymbol{z}_0 + \frac{\alpha_1}{4}(\boldsymbol{z}_1-\boldsymbol{z}_0) + \frac{\alpha_2}{4}(\boldsymbol{z}_2-\boldsymbol{z}_0)$ with $\alpha := (\alpha_1,\alpha_2)$ and $0 \le \alpha_1+\alpha_2 \le 4$. Let λ_i be the barycentric coordinate associated with the vertex $\boldsymbol{z}_i$ for all $i \in \{0,1,2\}$. Then the shape functions associated with the vertices are

$$\begin{aligned}\theta_{0,0}(\boldsymbol{x}) &= \frac{1}{3}\lambda_0(4\lambda_0-3)(2\lambda_0-1)(4\lambda_0-1),\\ \theta_{4,0}(\boldsymbol{x}) &= \frac{1}{3}\lambda_1(4\lambda_1-3)(2\lambda_1-1)(4\lambda_1-1),\\ \theta_{0,4}(\boldsymbol{x}) &= \frac{1}{3}\lambda_2(4\lambda_2-3)(2\lambda_2-1)(4\lambda_2-1).\end{aligned}$$

Those associated with the first edge (connecting $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$) are

$$\begin{aligned}\theta_{3,1}(\boldsymbol{x}) &= \frac{16}{3}\lambda_1\lambda_2(2\lambda_1-1)(4\lambda_1-1),\\ \theta_{2,2}(\boldsymbol{x}) &= 4\lambda_1\lambda_2(4\lambda_1-1)(4\lambda_2-1),\\ \theta_{1,3}(\boldsymbol{x}) &= \frac{16}{3}\lambda_1\lambda_2(2\lambda_2-1)(4\lambda_2-1).\end{aligned}$$

Those associated with the second edge (connecting $\boldsymbol{z}_2$ to $\boldsymbol{z}_0$) are

$$\begin{aligned}\theta_{0,3}(\boldsymbol{x}) &= \frac{16}{3}\lambda_0\lambda_2(2\lambda_2-1)(4\lambda_2-1),\\ \theta_{0,2}(\boldsymbol{x}) &= 4\lambda_0\lambda_2(4\lambda_0-1)(4\lambda_2-1),\\ \theta_{0,1}(\boldsymbol{x}) &= \frac{16}{3}\lambda_0\lambda_2(2\lambda_0-1)(4\lambda_0-1).\end{aligned}$$

Those associated with the third edge (connecting $\boldsymbol{z}_0$ to $\boldsymbol{z}_1$) are

$$\begin{aligned}\theta_{1,0}(\boldsymbol{x}) &= \frac{16}{3}\lambda_0\lambda_1(2\lambda_0-1)(4\lambda_0-1),\\ \theta_{2,0}(\boldsymbol{x}) &= 4\lambda_0\lambda_1(4\lambda_0-1)(4\lambda_1-1),\\ \theta_{3,0}(\boldsymbol{x}) &= \frac{16}{3}\lambda_0\lambda_1(2\lambda_1-1)(4\lambda_1-1).\end{aligned}$$

Finally, those associated with the three internal Lagrange nodes are

$$\begin{aligned}\theta_{1,1}(\boldsymbol{x}) &= 32\lambda_0\lambda_1\lambda_2(4\lambda_0-1),\\ \theta_{2,1}(\boldsymbol{x}) &= 32\lambda_0\lambda_1\lambda_2(4\lambda_1-1),\\ \theta_{1,2}(\boldsymbol{x}) &= 32\lambda_0\lambda_1\lambda_2(4\lambda_2-1).\end{aligned}$$

Exercise 7.10 (Quadratic Crouzeix–Raviart). (i) We observe that

$$\lambda_0(\boldsymbol{g}_j)^2 + \lambda_1(\boldsymbol{g}_j)^2 + \lambda_2(\boldsymbol{g}_j)^2 = \alpha^2 + (1-\alpha)^2, \qquad \forall j \in \{1{:}6\}.$$

(ii) Since the nonzero polynomial $q(\boldsymbol{x}) = \lambda_0(\boldsymbol{x})^2 + \lambda_1(\boldsymbol{x})^2 + \lambda_2(\boldsymbol{x})^2 - \alpha^2 + (1-\alpha)^2$ is such that $\boldsymbol{\sigma}_j(q) = 0$ for all $j \in \{1{:}6\}$, we conclude that the unisolvence property does not hold. Hence, $(K, \mathbb{P}_{2,2}, \Sigma)$ is not a finite element.
(iii) Let $\xi_0 := -\frac{\sqrt{3}}{3}$, $\xi_1 := \frac{\sqrt{3}}{3}$ be the two nodes of the Gauss–Legendre quadrature of order 3 and let $\omega_0 = \omega_1 := 1$ be the corresponding weights (see Table 6.1). Let $\boldsymbol{a}_{2i+k} := \boldsymbol{T}_{F_i}(\xi_k)$. Setting $\alpha := \frac{1}{2}(\xi_0+1)$, we have $\boldsymbol{a}_4 = (\alpha, 0)$, $\boldsymbol{a}_5 = (1-\alpha, 0)$, $\boldsymbol{a}_0 = (1-\alpha, \alpha)$, $\boldsymbol{a}_1 = (\alpha, 1-\alpha)$, $\boldsymbol{a}_3 = (0, 1-\alpha)$, $\boldsymbol{a}_2 = (0, \alpha)$. Let $p(\boldsymbol{x}) := \lambda_0(\boldsymbol{x})^2 + \lambda_1(\boldsymbol{x})^2 + \lambda_2(\boldsymbol{x})^2 - \alpha^2 + (1-\alpha)^2$. From Step (ii), we infer that $p(\boldsymbol{a}_j) = 0$ for all $j \in \{0{:}5\}$. Since the quadrature is of order three, this shows that

$$\varpi_{2i+k}(p) = \frac{1}{2}(\omega_0 p(\boldsymbol{a}_{2i+0}) + \omega_1 p(\boldsymbol{a}_{2i+1})) = 0,$$

for all $i \in \{0{:}2\}$ and all $k \in \{0{:}1\}$. Hence, the triple $(K, \mathbb{P}_{2,2}, \Sigma)$ is not a finite element.

Chapter 8

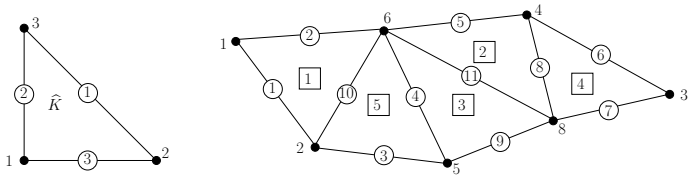
Meshes

Exercises

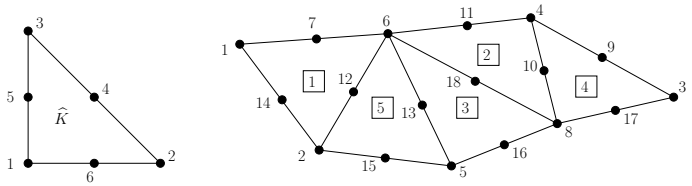
Exercise 8.1 (Curved triangle). Consider the $\mathbb{P}_2$ transformation of a triangle shown in the upper right panel of Figure 8.1. Consider a geometric node of K that is the image of the midpoint of an edge of $\widehat{K}$. Show that the tangent vector to the curved boundary at this node is collinear to the vector formed by the two vertices of the corresponding curved edge. (*Hint*: use the properties of the Lagrange $\mathbb{P}_2$ shape functions.)

Exercise 8.2 (Euler relations). Let $\mathcal{T}_h$ be a matching mesh in $\mathbb{R}^2$ composed of polygons all having ν vertices. (i) Show that $2N_\mathrm{e} - N_\mathrm{e}^\partial = \nu N_\mathrm{c}$. (ii) Combine this result with the Euler relations to show that $N_\mathrm{c} \sim \frac{2}{\nu-2} N_\mathrm{v}$ and $N_\mathrm{e} \sim \frac{\nu}{\nu-2} N_\mathrm{v}$ for fine enough meshes where $N_\mathrm{v}^\partial = N_\mathrm{e}^\partial \ll \min(N_\mathrm{v}, N_\mathrm{e}, N_\mathrm{c})$.

Exercise 8.3 (Connectivity arrays `j_cv`, `j_ce`). Write admissible connectivity arrays `j_cv` and `j_ce` for the following mesh where the face enumeration is identified with large circles and the cell enumeration with squares.



Exercise 8.4 (Connectivity array `j_geo`). Define a connectivity array `j_geo` for the following mesh such that the determinant of the Jacobian matrix of $\boldsymbol{T}_K$ is positive for all the cells.



Exercise 8.5 (Geometric mapping). Let $\boldsymbol{z}_1 := (0,0)$, $\boldsymbol{z}_2 := (1,0)$, $\boldsymbol{z}_3 := (0,1)$, $\boldsymbol{z}_4 := (\frac{1}{3}, \frac{1}{3})$. Consider the triangles $K_1 := \mathrm{conv}(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_4)$, $K_2 := \mathrm{conv}(\boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4)$, and $K_3 := \mathrm{conv}(\boldsymbol{z}_3, \boldsymbol{z}_1, \boldsymbol{z}_4)$.

(i) Construct the affine geometric mappings $\boldsymbol{T}_{K_2} : K_1 \to K_2$ and $\boldsymbol{T}_{K_3} : K_1 \to K_3$ s.t. $\boldsymbol{T}_{K_2}(\boldsymbol{z}_1) = \boldsymbol{z}_2$, $\boldsymbol{T}_{K_2}(\boldsymbol{z}_4) = \boldsymbol{z}_4$, and $\boldsymbol{T}_{K_3}(\boldsymbol{z}_1) = \boldsymbol{z}_3$, $\boldsymbol{T}_{K_3}(\boldsymbol{z}_4) = \boldsymbol{z}_4$. (*Hint*: $\boldsymbol{T}_{K_2}$ is of the form $\boldsymbol{T}_{K_2}(\boldsymbol{x}) = \boldsymbol{z}_2 + \mathbb{J}_{K_2}(\boldsymbol{x} - \boldsymbol{z}_1)$.) (ii) Compute $\det(\mathbb{J}_{K_2})\mathbb{J}_{K_2}^{-1}$ and $\det(\mathbb{J}_{K_3})\mathbb{J}_{K_3}^{-1}$. *Note*: the transformation $\boldsymbol{v} \mapsto \det(\mathbb{J}_K)\mathbb{J}_K^{-1}\boldsymbol{v}\circ\boldsymbol{T}_K$ is called contravariant Piola transformation; see (9.9c).

Solution to exercises

Exercise 8.1 (Curved triangle). To fix the ideas, consider the enumeration of the nodes in $\widehat{K}$ as depicted in the central panel of Figure 8.2, and consider the tangent vector to the boundary of K at the node $\boldsymbol{a}_6 = \boldsymbol{T}_K(\widehat{\boldsymbol{a}}_6)$. This tangent vector is collinear to $\sum_{i=1}^6 \boldsymbol{a}_i \partial_{\widehat{x}_1}\widehat{\psi}_i(\widehat{\boldsymbol{a}}_6)$, and owing to the properties of the shape functions of the Lagrange $\mathbb{P}_2$ finite element, we infer that $-\partial_{\widehat{x}_1}\widehat{\psi}_1(\widehat{a}_6) = \partial_{\widehat{x}_1}\widehat{\psi}_2(\widehat{a}_6) \neq 0$, whereas we have $\partial_{\widehat{x}_1}\widehat{\psi}_i(\widehat{a}_6) = 0$ for all $i \in \{3,4,5,6\}$. Hence, the tangent vector is colinear to $\partial_{\widehat{x}_1}\widehat{\psi}_2(\widehat{a}_6)(\boldsymbol{a}_2 - \boldsymbol{a}_1)$.

Exercise 8.2 (Euler relations). (i) Separating all the mesh cells, we obtain νN_c edges (since the boundary of each polygon consists of ν faces), and this number is equal to $2N_\mathrm{e} - N_\mathrm{e}^\partial$ since each edge leads to two edges except the boundary edges, i.e., $\nu N_\mathrm{c} = 2N_\mathrm{e} - N_\mathrm{e}^\partial$.
(ii) Combined with the Euler relations $N_\mathrm{c} = N_\mathrm{e} - N_\mathrm{v} + 1 - I$ and $N_\mathrm{e}^\partial = N_\mathrm{v}^\partial$, we infer that $\frac{\nu-2}{2}N_\mathrm{c} = N_\mathrm{v} - \frac{1}{2}N_\mathrm{v}^\partial + I - 1$, so that $N_\mathrm{c} \sim \frac{2}{\nu-2}N_\mathrm{v}$ for fine meshes. Finally, $N_\mathrm{e} \sim \frac{\nu}{2}N_\mathrm{c} \sim \frac{\nu}{\nu-2}N_\mathrm{v}$.

Exercise 8.3 (Connectivity arrays `j_cv`, `j_ce`). One possibility is the following connectivity arrays:

$$\texttt{j\_cv} = \begin{pmatrix} 1 & 2 & 6 \\ 6 & 8 & 4 \\ 5 & 8 & 6 \\ 4 & 8 & 3 \\ 2 & 5 & 6 \end{pmatrix}, \qquad \texttt{j\_ce} = \begin{pmatrix} 10 & 2 & 1 \\ 8 & 5 & 11 \\ 11 & 4 & 9 \\ 7 & 6 & 8 \\ 4 & 10 & 3 \end{pmatrix}.$$

Any permutation of the indices in each line is also legitimate, provided the same permutation is applied to `j_cv` and `j_ce`.

Exercise 8.4 (Connectivity array `j_geo`). The following array `j_geo` is such that the determinant of $\boldsymbol{T}_K$ is positive for all the cells:

$$\texttt{j\_geo} = \begin{pmatrix} 1 & 2 & 6 & 12 & 7 & 14 \\ 6 & 8 & 4 & 10 & 11 & 18 \\ 5 & 8 & 6 & 18 & 13 & 16 \\ 4 & 8 & 3 & 17 & 9 & 10 \\ 2 & 5 & 6 & 13 & 12 & 15 \end{pmatrix}.$$

It is possible to apply in each line any cyclic permutation to the indices of the first three columns and applying the same permutation to the indices of the last three columns.

Exercise 8.5 (Geometric mapping). (i) The geometric mapping $\boldsymbol{T}_{K_2}$ is necessarily of the form

$$\boldsymbol{T}_{K_2}(\boldsymbol{x}) = \boldsymbol{z}_2 + \mathbb{J}_{K_2}(\boldsymbol{x} - \boldsymbol{z}_1), \qquad \mathbb{J}_{K_2} := \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

since it is affine and $\boldsymbol{T}_{K_2}(\boldsymbol{z}_1) = \boldsymbol{z}_2$. The requirements $\boldsymbol{T}_{K_2}(\boldsymbol{z}_2) = \boldsymbol{z}_3$ and $\boldsymbol{T}_{K_2}(\boldsymbol{z}_4) = \boldsymbol{z}_4$ (this is the only possibility since $\boldsymbol{T}_{K_2}$ maps the vertices of K_1 to the vertices of K_2) lead to

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 1 & \frac{1}{3} \\ 0 & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{3} \\ 1 & \frac{1}{3} \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -1 & -\frac{2}{3} \\ 1 & \frac{1}{3} \end{pmatrix}.$$

Hence, we obtain

$$\mathbb{J}_{K_2} = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}.$$

Similarly, let us set

$$\boldsymbol{T}_{K_3}(\boldsymbol{x}) = \boldsymbol{z}_3 + \mathbb{J}_{K_3}(\boldsymbol{x} - \boldsymbol{z}_1), \qquad \mathbb{J}_{K_3} := \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Observing that $\boldsymbol{T}_{K_3}(\boldsymbol{z}_2) = \boldsymbol{z}_1$ and $\boldsymbol{T}_{K_3}(\boldsymbol{z}_4) = \boldsymbol{z}_4$, we infer that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{3} \\ 0 & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{3} \\ 0 & \frac{1}{3} \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{3} \\ -1 & -\frac{2}{3} \end{pmatrix}.$$

Hence, we obtain

$$\mathbb{J}_{K_3} = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}.$$

(ii) We have $\det(\mathbb{J}_{K_2}) = 1$ and $\mathbb{J}_{K_2}^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}$, so that

$$\det(\mathbb{J}_{K_2})\mathbb{J}_{K_2}^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} v \\ -u - v \end{pmatrix}.$$

Similarly, we have $\det(\mathbb{J}_{K_3}) = 1$ and $\mathbb{J}_{K_3}^{-1} = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}$, so that

$$\det(\mathbb{J}_{K_3})\mathbb{J}_{K_3}^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -u - v \\ u \end{pmatrix}.$$

Chapter 9

Finite element generation

Exercises

Exercise 9.1 (Canonical hybrid element). Consider an affine geometric mapping $\boldsymbol{T}_K$ and the pullback by $\boldsymbol{T}_K$ for ψ_K. Let $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ be the canonical hybrid element of §7.6. Verify that Proposition 9.2 generates the canonical hybrid element in K. Write the dofs.

Exercise 9.2 (Line measure). (i) Prove Lemma 9.12 for line measures. (*Hint*: the change in line measure is $\frac{\mathrm{d}l}{\mathrm{d}\widehat{l}}(\boldsymbol{x}) = \lim_{h\to 0} \frac{\|\boldsymbol{T}_K(\widehat{\boldsymbol{x}}+h\widehat{\boldsymbol{\tau}})-\boldsymbol{T}_K(\widehat{\boldsymbol{x}})\|_{\ell^2}}{\|h\widehat{\boldsymbol{\tau}}\|_{\ell^2}}$.) (ii) Assume that $d = 2$. Show that $|\det(\mathbb{J}_K)|\|\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}\|_{\ell^2(\mathbb{R}^2)} = \|\mathbb{J}_K\widehat{\boldsymbol{\tau}}\|_{\ell^2(\mathbb{R}^2)}$ for any pair of unit vectors $(\widehat{\boldsymbol{n}}, \widehat{\boldsymbol{\tau}})$ that are orthogonal.

Exercise 9.3 (Surface measure). (i) Let $\boldsymbol{T}_F := \boldsymbol{T}_{K|\widehat{F}} : \widehat{F} \to F$ and $\widehat{\boldsymbol{x}} \in \widehat{F}$. Let $\mathbb{J}_F(\widehat{\boldsymbol{x}}) \in \mathbb{R}^{d\times(d-1)}$ be the Jacobian matrix representing the (Fréchet) derivative $D\boldsymbol{T}_F(\widehat{\boldsymbol{x}})$. Let $\mathrm{g}_F(\widehat{\boldsymbol{x}}) = (\mathbb{J}_F(\widehat{\boldsymbol{x}}))^{\mathsf{T}}\mathbb{J}_F(\widehat{\boldsymbol{x}}) \in \mathbb{R}^{(d-1)\times(d-1)}$ be the surface metric tensor at $\widehat{\boldsymbol{x}}$. Prove that $\sqrt{\det(\mathrm{g}_F(\widehat{\boldsymbol{x}}))} = |\det(\mathbb{J}_K)|\,\|\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}\|_{\ell^2}$. (*Hint*: use that $\mathrm{d}s = \sqrt{\det(\mathrm{g}_F(\widehat{\boldsymbol{x}}))}\,\mathrm{d}\widehat{s}$.) (ii) Let $\widehat{K} := \{(\widehat{x}_1, \widehat{x}_2, \widehat{x}_3) \in \mathbb{R}^3 \mid 0 \le \widehat{x}_1, \widehat{x}_2, \widehat{x}_3,\ \widehat{x}_1 + \widehat{x}_2 + \widehat{x}_3 \le 1\}$ be the unit simplex in $\mathbb{R}^3$. Let $\boldsymbol{T}_K(\widehat{\boldsymbol{x}}) := (\widehat{x}_1, \widehat{x}_2, \widehat{x}_1^2 + \widehat{x}_2^2 - \widehat{x}_3)^{\mathsf{T}}$. Let $\widehat{F}$ be the face $\{\widehat{x}_3 = 0\}$ and $F := \boldsymbol{T}_K(\widehat{F})$. Compute $\mathbb{J}_F$, $\mathbb{J}_K$, g_F and verify the identity proved in Step (i).

Exercise 9.4 (Sobolev spaces). Prove that ψ_K^{g} is a bounded isomorphism from $H^1(K)$ to $H^1(\widehat{K})$, that $\boldsymbol{\psi}_K^{\mathrm{c}}$ is a bounded isomorphism from $\boldsymbol{H}(\mathrm{curl}; K)$ to $\boldsymbol{H}(\mathrm{curl}; \widehat{K})$, and that $\boldsymbol{\psi}_K^{\mathrm{d}}$ is a bounded isomorphism from $\boldsymbol{H}(\mathrm{div}; K)$ to $\boldsymbol{H}(\mathrm{div}; \widehat{K})$.

Exercise 9.5 (Transformation of cross products). Let $\mathbb{A}$ be a 3×3 invertible matrix. Prove that $\mathbb{A}^{-\mathsf{T}}(\boldsymbol{x}\times\boldsymbol{y}) = \det(\mathbb{A})^{-1}(\mathbb{A}\boldsymbol{x}\times\mathbb{A}\boldsymbol{y})$ for any vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3$.

Exercise 9.6 ((9.15b)). Prove (9.15b).

Solution to exercises

Exercise 9.1 (Canonical hybrid element). Let $\{\mu_m\}_{m\in\{1:n_{\mathrm{sh}}^{\mathrm{e}}\}}$ be a basis of $\mathbb{P}_{k-2,1}$, $\{\zeta_m\}_{m\in\{1:n_{\mathrm{sh}}^{\mathrm{f}}\}}$ be a basis of $\mathbb{P}_{k-3,2}$, and $\{\psi_m\}_{m\in\{1:n_{\mathrm{sh}}^{\mathrm{c}}\}}$ be a basis of $\mathbb{P}_{k-4,3}$. Let $K = \boldsymbol{T}_K(\widehat{K})$. For the vertex

dofs, we have with $\boldsymbol{z} = \boldsymbol{T}_K(\widehat{\boldsymbol{z}})$,

$$\sigma_{\boldsymbol{z}}^{\mathrm{v}}(v) = \widehat{\sigma}_{\widehat{\boldsymbol{z}}}^{\mathrm{v}}(v \circ \boldsymbol{T}_K) = (v \circ \boldsymbol{T}_K)(\widehat{\boldsymbol{z}}) = v(\boldsymbol{z}),$$

for all $i \in \{0{:}d\}$. For the edge dofs, letting $E = \boldsymbol{T}_K(\widehat{E})$, we observe that

$$\begin{aligned}\sigma_{E,m}^{\mathrm{e}}(v) = \widehat{\sigma}_{\widehat{E},m}^{\mathrm{e}}(v \circ \boldsymbol{T}_K) &= \frac{1}{|\widehat{E}|}\int_{\widehat{E}} (v \circ \boldsymbol{T}_K)(\mu_m \circ \boldsymbol{T}_{\widehat{E}}^{-1})\,\mathrm{d}\widehat{l} \\ &= \frac{1}{|\widehat{E}|}\int_{\widehat{E}} (v \circ \boldsymbol{T}_K)((\mu_m \circ \boldsymbol{T}_{\widehat{E}}^{-1} \circ \boldsymbol{T}_K^{-1}) \circ \boldsymbol{T}_K)\,\mathrm{d}\widehat{l} \\ &= \frac{1}{|E|}\int_E v(\mu_m \circ \boldsymbol{T}_{K,E}^{-1})\,\mathrm{d}l,\end{aligned}$$

for all $m \in \{1{:}n_{\mathrm{sh}}^{\mathrm{e}}\}$, where $\boldsymbol{T}_{K,E} \coloneqq \boldsymbol{T}_K \circ \boldsymbol{T}_{\widehat{E}}$ maps $\mathbb{R}$ to the line in $\mathbb{R}^3$ supporting the edge E. Proceeding similarly for the face dofs and setting $F \coloneqq \boldsymbol{T}_K(\widehat{F})$, we infer that

$$\sigma_{F,m}^{\mathrm{f}}(v) = \frac{1}{|F|}\int_F p(\zeta_m \circ \boldsymbol{T}_{K,F}^{-1})\,\mathrm{d}s,$$

for all $m \in \{1{:}n_{\mathrm{sh}}^{\mathrm{f}}\}$, where we have set $\boldsymbol{T}_{K,F} \coloneqq \boldsymbol{T}_K \circ \boldsymbol{T}_{\widehat{F}}$ which maps $\mathbb{R}^2$ to the plane in $\mathbb{R}^3$ supporting the face F. For the cell dofs, we finally find that

$$\sigma_m^{\mathrm{c}}(p) = \frac{1}{|K|}\int_K p(\psi_m \circ \boldsymbol{T}_K^{-1})\,\mathrm{d}x,$$

for all $m \in \{1{:}n_{\mathrm{sh}}^{\mathrm{c}}\}$.

Exercise 9.2 (Line measure). (i) Let $\widehat{E}$ be an edge of $\widehat{K}$ and let $\widehat{\boldsymbol{x}}$ be a point in the interior of $\widehat{E}$. There is no ambiguity to define a unit vector tangent to $\widehat{E}$ at $\widehat{\boldsymbol{x}}$, say $\widehat{\boldsymbol{\tau}}$ (note that there are two choices for $\widehat{\boldsymbol{\tau}}$). Let $E \coloneqq \boldsymbol{T}_K(\widehat{E})$. The change in line measure between $\widehat{E}$ and E is by definition

$$\frac{\mathrm{d}l}{\mathrm{d}\widehat{l}} = \lim_{h\to 0}\frac{\|\boldsymbol{T}_K(\widehat{\boldsymbol{x}} + h\widehat{\boldsymbol{\tau}}) - \boldsymbol{T}_K(\widehat{\boldsymbol{x}})\|_{\ell^2}}{\|h\widehat{\boldsymbol{\tau}}\|_{\ell^2}}.$$

Using the definition of the Fréchet derivative, we have

$$\begin{aligned}\left|\frac{\mathrm{d}l}{\mathrm{d}\widehat{l}} - \|\mathbb{J}_K\widehat{\boldsymbol{\tau}}\|_{\ell^2}\right| &= \left|\frac{\mathrm{d}l}{\mathrm{d}\widehat{l}} - \lim_{h\to 0}\frac{\|D\boldsymbol{T}_K(\widehat{\boldsymbol{x}})(h\widehat{\boldsymbol{\tau}})\|_{\ell^2}}{\|h\widehat{\boldsymbol{\tau}}\|_{\ell^2}}\right| \\ &\le \left|\lim_{h\to 0}\frac{\|\boldsymbol{T}_K(\widehat{\boldsymbol{x}} + h\widehat{\boldsymbol{\tau}}) - \boldsymbol{T}_K(\widehat{\boldsymbol{x}}) - D\boldsymbol{T}_K(\widehat{\boldsymbol{x}})(h\widehat{\boldsymbol{\tau}})\|_{\ell^2}}{\|h\widehat{\boldsymbol{\tau}}\|_{\ell^2}}\right| = 0.\end{aligned}$$

Hence, $\mathrm{d}l = \|\mathbb{J}_K\widehat{\boldsymbol{\tau}}\|_{\ell^2}\,\mathrm{d}\widehat{l}$.
(ii) When $d = 2$, the two statements in Lemma 9.12 are identical. Indeed in this case, $\mathrm{d}s = \mathrm{d}l$ and $\mathrm{d}\widehat{s} = \mathrm{d}\widehat{l}$. Hence, $|\det(\mathbb{J}_K)|\,\|\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}\|_{\ell^2(\mathbb{R}^2)} = \|\mathbb{J}_K\widehat{\boldsymbol{\tau}}\|_{\ell^2(\mathbb{R}^2)}$. Another (longer) way to proceed consists of using the cofactor formula for $\mathbb{J}_K^{-\mathsf{T}}$ and do the full computation.

Exercise 9.3 (Surface measure). (i) This is a simple consequence of Lemma 9.12. Since $\mathrm{d}s = |\det(\mathbb{J}_K)|\,\|\mathbb{J}_K^{-\mathsf{T}}\|_{\ell^2}\,\mathrm{d}\widehat{s}$, we infer that indeed

$$\sqrt{\det(\mathrm{g}_F(\widehat{\boldsymbol{x}}))} = |\det(\mathbb{J}_K)|\|\mathbb{J}_K^{-\mathsf{T}}\|_{\ell^2}.$$

(ii) Using the definition of $\boldsymbol{T}_K$, we have

$$\mathbb{J}_F = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2\widehat{x}_1 & 2\widehat{x}_2 \end{bmatrix}, \quad \mathrm{g}_F = \begin{bmatrix} 1+4\widehat{x}_1^2 & 4\widehat{x}_1\widehat{x}_2 \\ 4\widehat{x}_1\widehat{x}_2 & 1+4\widehat{x}_2^2 \end{bmatrix},$$

and

$$\mathbb{J}_K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2\widehat{x}_1 & 2\widehat{x}_2 & -1 \end{bmatrix}, \quad \det(\mathbb{J}_K) = -1, \quad \mathbb{J}_K^{-\mathsf{T}} = \begin{bmatrix} 1 & 0 & 2\widehat{x}_1 \\ 0 & 1 & 2\widehat{x}_2 \\ 0 & 0 & -1 \end{bmatrix}.$$

By definition, $\widehat{\boldsymbol{n}}_{\widehat{F}} = (0,0,-1)^\mathsf{T}$, so that $\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{F}} = (-2\widehat{x}_1, -2\widehat{x}_2, 1)^\mathsf{T}$. We conclude that

$$\sqrt{\det(\mathrm{g}_F(\widehat{\boldsymbol{x}}))} = \sqrt{1+4\widehat{x}_1^2+4\widehat{x}_2^2} = |\det(\mathbb{J}_K)|\,\|\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{F}}\|_{\ell^2}.$$

Exercise 9.4 (Sobolev spaces). The assertions are direct consequences of Lemma 9.6 and the fact that the geometric mapping $\boldsymbol{T}_K$ has bounded derivatives of any order.

Exercise 9.5 (Transformation of cross products). The key is to remember that the mixed product of three vectors $\boldsymbol{z}\cdot(\boldsymbol{x}\times\boldsymbol{y})$ is equal to $\det[\boldsymbol{z},\boldsymbol{x},\boldsymbol{y}]$ where $[\boldsymbol{z},\boldsymbol{x},\boldsymbol{y}]$ is the matrix with column vectors $\boldsymbol{z},\boldsymbol{x},\boldsymbol{y}$. We have

$$\begin{aligned}\boldsymbol{z}^\mathsf{T}\mathbb{A}^{-\mathsf{T}}(\boldsymbol{x}\times\boldsymbol{y}) &= (\mathbb{A}^{-1}\boldsymbol{z})\cdot(\boldsymbol{x}\times\boldsymbol{y}) = \det[\mathbb{A}^{-1}\boldsymbol{z},\boldsymbol{x},\boldsymbol{y}] = \det(\mathbb{A}^{-1}[\boldsymbol{z},\mathbb{A}\boldsymbol{x},\mathbb{A}\boldsymbol{y}]) \\ &= \det(\mathbb{A}^{-1})\det[\boldsymbol{z},\mathbb{A}\boldsymbol{x},\mathbb{A}\boldsymbol{y}] = \det(\mathbb{A})^{-1}\boldsymbol{z}^\mathsf{T}(\mathbb{A}\boldsymbol{x}\times\mathbb{A}\boldsymbol{y}).\end{aligned}$$

This proves the expected identity since $\boldsymbol{z}$ is arbitrary.

Exercise 9.6 ((9.15b)**).** The following holds true:

$$\begin{aligned}\int_E (\boldsymbol{v}\cdot\boldsymbol{\Phi}_K^{\mathrm{c}}(\widehat{\boldsymbol{\tau}}_{\widehat{E}}))(\boldsymbol{x})q(\boldsymbol{x})\,\mathrm{d}l &= \int_{\widehat{E}} (\boldsymbol{v}\cdot\boldsymbol{\Phi}_K^{\mathrm{c}}(\widehat{\boldsymbol{\tau}}_{\widehat{E}}))(\boldsymbol{T}_K(\widehat{\boldsymbol{x}}))\psi_K^{\mathrm{g}}(q)(\widehat{\boldsymbol{x}})\|\mathbb{J}_K\widehat{\boldsymbol{\tau}}_{\widehat{E}}\|_{\ell^2}\,\mathrm{d}\widehat{l} \\ &= \int_{\widehat{E}} ((\boldsymbol{v}\circ\boldsymbol{T}_K)\cdot(\mathbb{J}_K\widehat{\boldsymbol{\tau}}_{\widehat{E}}))(\widehat{\boldsymbol{x}})\psi_K^{\mathrm{g}}(q)(\widehat{\boldsymbol{x}})\,\mathrm{d}\widehat{l} \\ &= \int_{\widehat{E}} (\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{v})\cdot\widehat{\boldsymbol{\tau}}_{\widehat{E}})(\widehat{\boldsymbol{x}})\psi_K^{\mathrm{g}}(q)(\widehat{\boldsymbol{x}})\,\mathrm{d}\widehat{l}.\end{aligned}$$

Chapter 10

Mesh orientation

Exercises

Exercise 10.1 (Faces in 2D). Let $\boldsymbol{R}_{\frac{\pi}{2}}$ be the rotation of angle $\frac{\pi}{2}$ in $\mathbb{R}^2$. (i) Let $\mathbb{A}$ be an inversible 2×2 matrix. Prove that $\mathbb{A}^{-\mathsf{T}}\boldsymbol{R}_{\frac{\pi}{2}} = \frac{1}{\det(\mathbb{A})}\boldsymbol{R}_{\frac{\pi}{2}}\mathbb{A}$. (ii) Prove that $\boldsymbol{\Phi}_K^{\mathrm{d}}(\boldsymbol{R}_{\frac{\pi}{2}}(\boldsymbol{z})) = \boldsymbol{R}_{\frac{\pi}{2}}(\boldsymbol{\Phi}_K^{\mathrm{c}}(\boldsymbol{z}))$ for all $\boldsymbol{z}\in\mathbb{R}^2$.

Exercise 10.2 (Connectivity arrays `j_cv`, `j_ce`). Consider the mesh shown in Figure 10.1, where the face enumeration is identified with large circles and the cell enumeration is identified with squares. (i) Write the connectivity arrays `j_cv` and `j_ce` based on increasing vertex-index

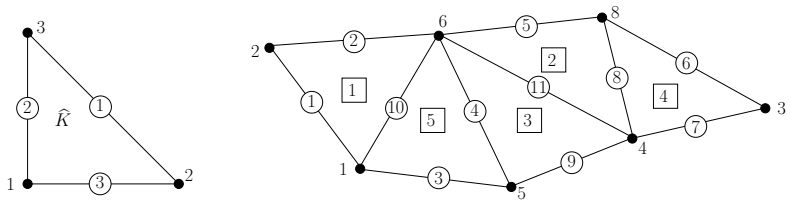


Figure 10.1: Illustration for Exercise 10.2.

enumeration. (ii) Give the sign of the determinant of the Jacobian matrix of $\boldsymbol{T}_K$ for each triangle.

Exercise 10.3 (Connectivity array `j_geo`). Consider the mesh shown in Figure 10.2 and based on the $\mathbb{P}_{2,2}$ geometric Lagrange element. (i) Write the connectivity array `j_geo` based on increasing vertex-index enumeration. (ii) Give the sign of the determinant of the Jacobian matrix of $\boldsymbol{T}_K$ for each triangle.

Exercise 10.4 (Orientation of quadrangular mesh). (i) Using the enumeration and the orientation conventions proposed in this chapter, orient the mesh shown in Figure 10.3, where the cell enumeration is identified with shaded rectangles. (ii) Give the connectivity array `j_geo` so that the mesh orientation is generation-compatible and the determinant of the Jacobian matrix of $\boldsymbol{T}_K$ is positive for even quadrangles and negative for odd quadrangles.

Exercise 10.5 (Mesh extrusion). (i) Let K be a triangular prism. Denote by $\boldsymbol{e}_3$ the unit vector in the vertical direction. Let $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3$ be the three vertices of the bottom triangular face of K,

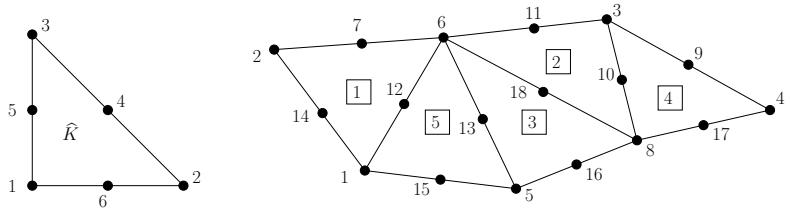


Figure 10.2: Illustration for Exercise 10.3.

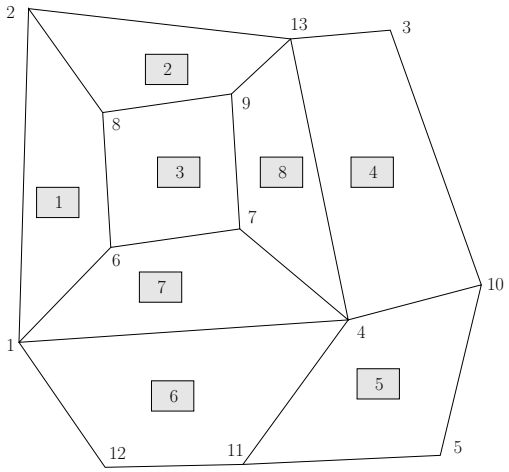


Figure 10.3: Illustration for Exercise 10.4.

and let $\boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6$ be the three vertices of its top triangular face, so that the segments $[\boldsymbol{z}_p, \boldsymbol{z}_{p+3}]$ are parallel to $\boldsymbol{e}_3$ for every $p \in \{1,2,3\}$. Propose a way to cut K into three tetrahedra. (ii) Let $\mathcal{T}_h$ be a two-dimensional oriented mesh composed of triangles. Let $\mathcal{T}_h'$ be a copy of $\mathcal{T}_h$ obtained by translating $\mathcal{T}_h$ in the third direction $\boldsymbol{e}_3$, say $\mathcal{T}_h' := \mathcal{T}_h + \boldsymbol{e}_3$. Propose a way to cut all the prisms thus formed to make a matching mesh composed of tetrahedra.

Solution to exercises

Exercise 10.1 (Faces in 2D). (i) Let us set $\mathbb{A} := \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. We have

$$\mathbb{A}^{-\mathsf{T}} \boldsymbol{R}_{\frac{\pi}{2}} = \frac{1}{\det(\mathbb{A})} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \frac{1}{\det(\mathbb{A})} \begin{pmatrix} -c & -d \\ a & b \end{pmatrix}.$$

We also have

$$\boldsymbol{R}_{\frac{\pi}{2}} \mathbb{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} -c & -d \\ a & b \end{pmatrix}.$$

This proves the claim.
(ii) Using the above result and the definition of $\boldsymbol{\Phi}_K^{\mathrm{d}}$ and $\boldsymbol{\Phi}_K^{\mathrm{c}}$, we obtain

$$\boldsymbol{\Phi}_K^{\mathrm{d}}(\boldsymbol{R}_{\frac{\pi}{2}}(\boldsymbol{z})) = \epsilon_K \frac{\mathbb{J}_K^{-\mathsf{T}} \boldsymbol{R}_{\frac{\pi}{2}}(\boldsymbol{z})}{\|\mathbb{J}_K^{-\mathsf{T}} \boldsymbol{R}_{\frac{\pi}{2}}(\boldsymbol{z})\|_{\ell^2}} = \epsilon_K \frac{|\det(\mathbb{J}_K)|}{\det(\mathbb{J}_K)} \frac{\boldsymbol{R}_{\frac{\pi}{2}}(\mathbb{J}_K \boldsymbol{z})}{\|\boldsymbol{R}_{\frac{\pi}{2}}(\mathbb{J}_K \boldsymbol{z})\|_{\ell^2}}$$
$$= \boldsymbol{R}_{\frac{\pi}{2}}\left(\frac{\mathbb{J}_K \boldsymbol{z}}{\|\mathbb{J}_K \boldsymbol{z}\|_{\ell^2}}\right) = \boldsymbol{R}_{\frac{\pi}{2}}(\boldsymbol{\Phi}_K^{\mathrm{c}}(\boldsymbol{z})).$$

Exercise 10.2 (Connectivity arrays `j_cv`, `j_ce`). (i) The connectivity arrays are

$$\texttt{j\_cv} = \begin{pmatrix} 1 & 2 & 6 \\ 4 & 6 & 8 \\ 4 & 5 & 6 \\ 3 & 4 & 8 \\ 1 & 5 & 6 \end{pmatrix}, \qquad \texttt{j\_ce} = \begin{pmatrix} 2 & 10 & 1 \\ 5 & 8 & 11 \\ 4 & 11 & 9 \\ 8 & 6 & 7 \\ 4 & 10 & 3 \end{pmatrix}.$$

(ii) The signs of the determinants are as follows:

$$\begin{bmatrix} \text{index of } K\text{:} & 1 & 2 & 3 & 4 & 5 \\ \text{sign}(\det \mathbb{J}_K)\text{:} & - & - & - & - & + \end{bmatrix}.$$

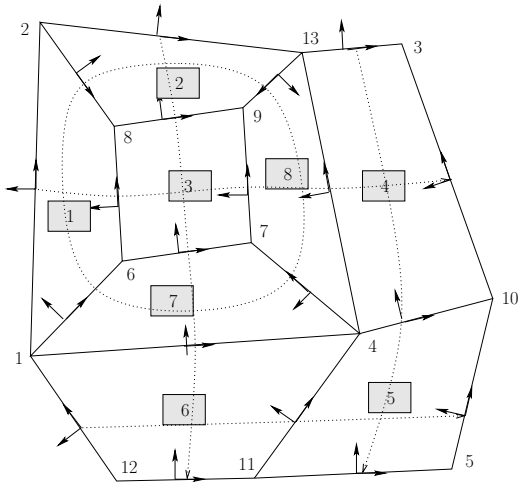
Exercise 10.3 (Connectivity array `j_geo`). (i) The following array `j_geo` is based on increasing vertex-index enumeration:

$$\texttt{j\_geo} = \begin{pmatrix} 1 & 2 & 6 & 7 & 12 & 14 \\ 3 & 6 & 8 & 18 & 10 & 11 \\ 5 & 6 & 8 & 18 & 16 & 13 \\ 3 & 4 & 8 & 17 & 10 & 9 \\ 1 & 5 & 6 & 13 & 12 & 15 \end{pmatrix}.$$

(ii) The signs of the determinants are as follows:

$$\begin{bmatrix} \text{index of } K\text{:} & 1 & 2 & 3 & 4 & 5 \\ \text{sign}(\det \mathbb{J}_K)\text{:} & - & + & - & - & + \end{bmatrix}.$$

Exercise 10.4 (Orientation of quadrangular mesh). (i) A generation-compatible orientation of the edges and faces is as follows:



The edges belonging to the same connected component of the edge/cell graph are linked by a dotted curve.
(ii) If one wishes that the determinant of the Jacobian matrix is positive for even quadrangles and negative for odd quadrangles, the geometric connectivity is as follows:

$$\texttt{j\_geo} = \begin{bmatrix} 1 & 2 & 6 & 8 \\ 2 & 8 & 13 & 9 \\ 6 & 8 & 7 & 9 \\ 4 & 10 & 13 & 3 \\ 11 & 4 & 5 & 10 \\ 12 & 11 & 1 & 4 \\ 1 & 6 & 4 & 7 \\ 4 & 13 & 7 & 9 \end{bmatrix}.$$

Note that for each cell K_m, $m \in \{1{:}6\}$, $\texttt{j\_geo}(m, 1)$ gives the index of the vertex that is the origin of K_m (such that the two edges sharing it are oriented away from it).

Exercise 10.5 (Mesh extrusion). (i) We first orient the edges of the bottom face using the increasing vertex-index enumeration. Then one needs to find a strategy to cut the three vertical faces. The key idea is to use the orientation of the edges of the bottom face. The cutting of the face whose vertices are $(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_4, \boldsymbol{z}_5)$ is done by connecting $\boldsymbol{z}_1$ with $\boldsymbol{z}_5$, i.e., the cut starts from $\boldsymbol{z}_1$ and is done along the vector $(\boldsymbol{z}_2 - \boldsymbol{z}_1) + \boldsymbol{e}_3$. The cutting of the face whose vertices are $(\boldsymbol{z}_1, \boldsymbol{z}_3, \boldsymbol{z}_4, \boldsymbol{z}_6)$ is done by connecting $\boldsymbol{z}_1$ with $\boldsymbol{z}_6$, i.e., the cut starts from $\boldsymbol{z}_1$ and is done along the vector $(\boldsymbol{z}_3 - \boldsymbol{z}_1) + \boldsymbol{e}_3$. The cutting of the face whose vertices are $(\boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_5, \boldsymbol{z}_6)$ is done by connecting $\boldsymbol{z}_2$ with $\boldsymbol{z}_6$, i.e., the cut starts from $\boldsymbol{z}_2$ and is done along the vector $(\boldsymbol{z}_3 - \boldsymbol{z}_2) + \boldsymbol{e}_3$. The proposed cutting produces three tetrahedra, with vertices $(\boldsymbol{z}_1, \boldsymbol{z}_4, \boldsymbol{z}_5, \boldsymbol{z}_6)$, $(\boldsymbol{z}_1; \boldsymbol{z}_2, \boldsymbol{z}_5, \boldsymbol{z}_6)$, and $(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_6)$.
(ii) The key idea is to use the orientation of the edges of $\mathcal{T}_h$ to do the cutting of the vertical faces of the prisms produced by translating $\mathcal{T}_h$ in the $\boldsymbol{e}_3$ direction. Let E be an edge of $\mathcal{T}_h$ with vertices $\boldsymbol{z}_p$, $\boldsymbol{z}_q$ and orientation vector $\boldsymbol{\tau}_E$, and assume that $\boldsymbol{z}_q - \boldsymbol{z}_p$ and $\boldsymbol{\tau}_E$ have the same orientation (notice that if $p < q$, then $\boldsymbol{z}_q - \boldsymbol{z}_p$ and $\boldsymbol{\tau}_E$ have the same orientation if the increasing vertex-index enumeration technique is used). Let $\boldsymbol{z}_r := \boldsymbol{z}_p + \boldsymbol{e}_3$ and $\boldsymbol{z}_s := \boldsymbol{z}_q + \boldsymbol{e}_3$. Then we cut the vertical face whose vertices are $(\boldsymbol{z}_p, \boldsymbol{z}_q, \boldsymbol{z}_r, \boldsymbol{z}_s)$ by connecting $\boldsymbol{z}_p$ with $\boldsymbol{z}_s$, i.e., the cut starts from $\boldsymbol{z}_p$ and is done along the vector $\boldsymbol{\tau}_E + \boldsymbol{e}_3$. Notice that for the two prisms sharing the same rectangular face, the proposed strategy provides for a unique way to cut the face in question. As a result, the mesh of tetrahedra thus formed is a matching mesh.

Chapter 11

Local interpolation on affine meshes

Exercises

Exercise 11.1 (High-order derivative). Let two integers $m, d \geq 2$. Consider the map $\Phi : \{1{:}d\}^m \ni \mathtt{j} \longmapsto (\Phi_1(\mathtt{j}), \dots, \Phi_d(\mathtt{j})) \in \mathbb{N}^d$, where $\Phi_i(\mathtt{j}) := \mathrm{card}\{k \in \{1{:}m\} \mid \mathtt{j}_k = i\}$ for all $i \in \{1{:}d\}$, so that $|\Phi(\mathtt{j})| = m$ by construction. Let $C_{m,d} := \max_{\alpha \in \mathbb{N}^d, |\alpha|=m} \mathrm{card}\{\mathtt{j} \in \{1{:}d\}^m \mid \Phi(\mathtt{j}) = \alpha\}$. Let v be a smooth (scalar-valued) function. (i) Show that

$$\|D^m v\|_{\mathcal{M}_m(\mathbb{R}^d,\dots,\mathbb{R}^d;\mathbb{R})} \leq C_{m,d}^{\frac{1}{2}} \left(\sum_{\alpha \in \mathbb{N}^d, |\alpha|=m} |\partial^\alpha v|^2 \right)^{\frac{1}{2}}.$$

(ii) Show that $C_{m,2} = \max_{0 \leq l \leq m} \binom{m}{l} = 2^m$. (iii) Evaluate $C_{m,3}$ and $m \in \{2, 3\}$. (iv) Show that $\sum_{\alpha \in \mathbb{N}^d, |\alpha|=m} |\partial^\alpha v| \leq \binom{d+m-1}{d-1} \|D^m v\|_{\mathcal{M}_m(\mathbb{R}^d,\dots,\mathbb{R}^d;\mathbb{R})}$.

Exercise 11.2 (Flat triangle). Let K be a triangle with vertices $(0, 0)$, $(1, 0)$ and $(-1, \epsilon)$ with $0 < \epsilon \ll 1$. Consider the function $v(x_1, x_2) := x_1^2$. Evaluate the $\mathbb{P}_1$ Lagrange interpolant $\mathcal{I}_K^{\mathrm{L}}(v)$ (see (9.7)) and show that $|v - \mathcal{I}_K^{\mathrm{L}}(v)|_{H^1(K)} \geq \epsilon^{-1} |v|_{H^2(K)}$. (*Hint*: use a direct calculation of $\mathcal{I}_K^{\mathrm{L}}(v)$.)

Exercise 11.3 (Barycentric coordinate). Let K be a simplex with barycentric coordinates $\{\lambda_i\}_{i \in \{0:d\}}$. Prove that $|\lambda_i|_{W^{1,\infty}(K)} \leq \rho_K^{-1}$ for all $i \in \{0{:}d\}$.

Exercise 11.4 (Bramble–Hilbert). Prove Corollary 11.11. (*Hint*: use the Bramble–Hilbert/Deny–Lions lemma.)

Exercise 11.5 (Taylor polynomial). Let K be a convex cell. Consider a Lagrange finite element of degree $k \geq 1$ with nodes $\{\boldsymbol{a}_i\}_{i \in \mathcal{N}}$ and associated shape functions $\{\theta_i\}_{i \in \mathcal{N}}$. Consider a sufficiently smooth function v. For all $\boldsymbol{x}, \boldsymbol{y} \in K$, consider the Taylor polynomial of order k and the exact

remainder defined as follows:

$$\mathbb{T}_k(\boldsymbol{x},\boldsymbol{y}) := v(\boldsymbol{x}) + Dv(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) + \ldots + \frac{1}{k!}D^k v(\boldsymbol{x})(\underbrace{\boldsymbol{y}-\boldsymbol{x},\ldots,\boldsymbol{y}-\boldsymbol{x}}_{k \text{ times}}),$$

$$R_k(v)(\boldsymbol{x},\boldsymbol{y}) := \frac{1}{(k+1)!}D^{k+1}v(\eta\boldsymbol{x}+(1-\eta)\boldsymbol{y})(\underbrace{\boldsymbol{y}-\boldsymbol{x},\ldots,\boldsymbol{y}-\boldsymbol{x}}_{(k+1) \text{ times}}),$$

so that $v(\boldsymbol{y}) = \mathbb{T}_k(\boldsymbol{x},\boldsymbol{y}) + R_k(v)(\boldsymbol{x},\boldsymbol{y})$ for some $\eta \in [0,1]$. (i) Prove that $v(\boldsymbol{x}) = \mathcal{I}_K^{\mathrm{L}}(v)(\boldsymbol{x}) - \sum_{i\in\mathcal{N}} R_k(v)(\boldsymbol{x},\boldsymbol{a}_i)\theta_i(\boldsymbol{x})$, where $\mathcal{I}_K^{\mathrm{L}}$ is the Lagrange interpolant defined in (9.7). (*Hint*: interpolate with respect to $\boldsymbol{y}$.) (ii) Prove that $D^m v(\boldsymbol{x}) = D^m(\mathcal{I}_K^{\mathrm{L}}(v))(\boldsymbol{x}) - \sum_{i\in\mathcal{N}} R_k(v)(\boldsymbol{x},\boldsymbol{a}_i)D^m\theta_i(\boldsymbol{x})$ for all $m \le k$. (*Hint*: proceed as in (i), take m derivatives with respect to $\boldsymbol{y}$ at $\boldsymbol{x}$, and observe that $v(\boldsymbol{x}) = \mathbb{T}_k(\boldsymbol{x},\boldsymbol{x})$.) (iii) Deduce that $|v - \mathcal{I}_K^{\mathrm{L}}(v)|_{W^{m,\infty}(K)} \le c\sigma_K^m h_K^{k+1-m}|v|_{W^{k+1,\infty}(K)}$ with $c := \frac{1}{(k+1)!}c_* h_K^m \sum_{i\in\mathcal{N}} |\widehat{\theta}_i|_{W^{m,\infty}(\widehat{K})}$, where c_* comes from (11.7b) with $s=m$ and $p=\infty$.

Exercise 11.6 (L^p-stability of Lagrange interpolant). Let $\alpha \in (0,1)$. Consider the Lagrange $\mathbb{P}_1$ shape functions $\theta_1(x) := 1-x$ and $\theta_2(x) := x$. Consider the sequence of continuous functions $\{u_n\}_{n\in\mathbb{N}\setminus\{0\}}$ defined over the interval $K := [0,1]$ as $u_n(x) := n^\alpha - 1$ if $0 \le x \le \frac{1}{n}$ and $u_n(x) := x^{-\alpha} - 1$ otherwise. (i) Prove that the sequence is uniformly bounded in $L^p(0,1)$ for all p such that $p\alpha < 1$. (ii) Compute $\mathcal{I}_K^{\mathrm{L}}(u_n)$. Is the operator $\mathcal{I}_K^{\mathrm{L}}$ stable in the L^p-norm? (iii) Is the operator $\mathcal{I}_K^{\mathrm{L}}$ stable in any L^r-norm with $r \in [1,\infty)$?

Exercise 11.7 (Norm scaling, $s \notin \mathbb{N}$). Complete the proof of Lemma 11.7 for the case $s \notin \mathbb{N}$. (*Hint*: use (2.6) with $s = m+\sigma$, $m := \lfloor s \rfloor$, $\sigma := s - m \in (0,1)$.)

Exercise 11.8 (Morrey's polynomial). Let U be a nonempty open set in $\mathbb{R}^d$. Let $k \in \mathbb{N}$ and $p \in [1,\infty]$. Let $u \in W^{k,p}(U)$. Show that there is a unique polynomial $q \in \mathbb{P}_{k,d}$ s.t. $\int_U \partial^\alpha(u-q)\,\mathrm{d}x = 0$ for all $\alpha \in \mathbb{N}^d$ of length at most k. (*Hint*: see the proof of Lemma 11.9 and also Morrey [34, Thm. 3.6.10].)

Exercise 11.9 (Fractional Sobolev norm). Let $r \in (0,1)$. Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be an shape-regular affine mesh sequence and let $\widehat{K}$ be the reference element. Let K be an affine cell in $\mathcal{T}_h$. Using the notation $\widehat{v} := v \circ \boldsymbol{T}_K$, show that there is c such that $\|\widehat{v}\|_{H^r(\widehat{K})} \le c h_K^{r-\frac{d}{2}}|v|_{H^r(K)}$ for all $v \in H^r(K)$ such that $\int_K v\,\mathrm{d}x = 0$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$. (*Hint*: use Lemma 3.26.)

Solution to exercises

Exercise 11.1 (High-order derivative). (i) Let $\boldsymbol{h}_1,\ldots,\boldsymbol{h}_m$ be vectors in $\mathbb{R}^d$ such that $\|\boldsymbol{h}_l\|_{\ell^2(\mathbb{R}^d)} = 1$ for all $l \in \{1\!:\!m\}$. Owing to the Cauchy–Schwarz inequality, we infer that

$$\begin{aligned} D^m v(\boldsymbol{h}_1,\ldots,\boldsymbol{h}_m) &= \sum_{\mathtt{j}_1\in\{1:d\}} \cdots \sum_{\mathtt{j}_m\in\{1:d\}} h_{1,\mathtt{j}_1}\ldots h_{m,\mathtt{j}_m}\partial^{\Phi(\mathtt{j})}v \\ &\le \left(\sum_{\mathtt{j}\in\{1:d\}^m} |\partial^{\Phi(\mathtt{j})}v|^2\right)^{\frac{1}{2}}. \end{aligned}$$

As a result, we have

$$\|D^m v\|_{\mathcal{M}_m(\mathbb{R}^d,\dots,\mathbb{R}^d;\mathbb{R})} \le \left(\sum_{|\alpha|=m}\sum_{\Phi(\mathtt{j})=\alpha} |\partial^\alpha v|^2\right)^{\frac12},$$

where $\alpha\in\mathbb{N}^d$ in the first summation and $\mathtt{j}\in\{1{:}d\}^m$ in the second one.
(ii) For $d=2$, $\alpha=(l,m-l)$ with $l\in\{0{:}m\}$ and $\mathrm{card}\{\mathtt{j}\in\{1,2\}^m \mid \Phi(\mathtt{j})=(l,m-l)\}=\binom{m}{l}$.
(iii) A direct calculation shows that $C_{2,3}=3$ (attained, e.g., for $(1,1,0)$) and $C_{3,3}=6$ (attained for $\alpha=(1,1,1)$).
(iv) Since $|\partial^\alpha v|\le\|D^m v\|_{\mathcal{M}_m(\mathbb{R}^d,\dots,\mathbb{R}^d;\mathbb{R})}$ for all $\alpha\in\mathbb{N}^d$ with $|\alpha|=m$, the expected bound can be obtained with $\check{C}_{l,d}:=\mathrm{card}\{\alpha\in\mathbb{N}^d\mid|\alpha|=m\}=\binom{d+m-1}{d-1}$.

Exercise 11.2 (Flat triangle). We obtain $\mathcal{I}_K^{\mathrm{L}}(v)=x_1+2\epsilon^{-1}x_2$, so that $|v-\mathcal{I}_K^{\mathrm{L}}(v)|_{H^1(K)}=\int_K((2x_1-1)^2+\frac{4}{\epsilon^2})\,\mathrm{d}x\ge\frac{4}{\epsilon^2}|K|$, but $|v|^2_{H^2(K)}=4|K|$ is uniformly bounded w.r.t. ϵ.

Exercise 11.3 (Barycentric coordinate). Let $i\in\{0{:}d\}$. Let F_i be the face of K where λ_i vanishes. Let $\boldsymbol{n}_i$ be the outward unit normal to F_i. Let B be the largest ball that can be inscribed into K and let ρ_K be the diameter of B. Let S_i be the point where B is touching F_i and let N_i be the point opposite to S_i on ∂B. We have

$$1\ge\lambda(N_i)=\lambda(N_i)-\lambda(S_i)=-\rho_K\boldsymbol{n}_i{\cdot}\nabla\lambda_i=\rho_K\|\nabla\lambda_i\|_{\ell^2(\mathbb{R}^d)},$$

since $\lambda_i(S_i)=0$, $S_i=P_i+\rho_K\boldsymbol{n}_i$, λ_i is affine, and $\nabla\lambda_i$ is collinear to $\boldsymbol{n}_i$.

Exercise 11.4 (Bramble–Hilbert). For all $\pi\in\mathbb{P}_{k,d}$, $|f(v)|=|f(v+\pi)|\le\|f\|_{(W^{k+1,p}(S))'}\|v+\pi\|_{W^{k+1,p}(S)}$, that is, $|f(v)|\le\|f\|_{(W^{k+1,p}(S))'}\inf_{\pi\in\mathbb{P}_{k,d}}\|v+\pi\|_{W^{k+1,p}(S)}$. Thus, the assertion follows from Lemma 11.9.

Exercise 11.5 (Taylor polynomial). (i) Starting from $v(\boldsymbol{y})=\mathbb{T}_k(\boldsymbol{x},\boldsymbol{y})+R_k(v)(\boldsymbol{x},\boldsymbol{y})$ and interpolating with respect to $\boldsymbol{y}$ at any fixed $\boldsymbol{x}\in K$ leads to

$$\mathcal{I}_K^{\mathrm{L}}(v)(\boldsymbol{y})=\mathbb{T}_k(\boldsymbol{x},\boldsymbol{y})+\sum_{i\in\mathcal{N}}R_k(v)(\boldsymbol{x},\boldsymbol{a}_i)\theta_i(\boldsymbol{y}),$$

since the polynomial $\mathbb{T}_k(\boldsymbol{x},\boldsymbol{y})$ in $\boldsymbol{y}$ at fixed $\boldsymbol{x}$ is preserved by $\mathcal{I}_K^{\mathrm{L}}$. Evaluating the above expression at $\boldsymbol{y}=\boldsymbol{x}$ yields the assertion since $\mathbb{T}_k(\boldsymbol{x},\boldsymbol{x})=v(\boldsymbol{x})$.
(ii) Differentiating m times, $m\le k$, the above expression with respect to $\boldsymbol{y}$ at fixed $\boldsymbol{x}$ leads to

$$D^m(\mathcal{I}_K^{\mathrm{L}}(v))(\boldsymbol{y})=D^m\mathbb{T}_k(\boldsymbol{x},\boldsymbol{y})+\sum_{i\in\mathcal{N}}R_k(v)(\boldsymbol{x},\boldsymbol{a}_i)D^m\theta_i(\boldsymbol{y}),$$

and evaluating the expression at $\boldsymbol{y}=\boldsymbol{x}$ yields the assertion since $D^m\mathbb{T}_k(\boldsymbol{x},\boldsymbol{x})=D^m v(\boldsymbol{x})$.
(iii) Use the result of Step (ii) together with the triangle inequality and the bound $|R_k(v)(\boldsymbol{x},\boldsymbol{a}_i)|\le\frac{1}{(k+1)!}h_K^{k+1}|v|_{W^{k+1,\infty}(K)}$ for all $\boldsymbol{x}\in K$.

Exercise 11.6 (L^p-stability of Lagrange interpolant). (i) Observe that $\|u_n\|_{L^p(0,1)}\le\|x^{-\alpha}-1\|_{L^p(0,1)}\le1+\|x^{-\alpha}\|_{L^p(0,1)}\le1+\frac{1}{(1-p\alpha)^{\frac1p}}<\infty$ since $p\alpha<1$.
(ii) $\mathcal{I}_K^{\mathrm{L}}u_n(x)=u_n(0)\theta_1(x)+u_n(1)\theta_2(x)=(n^\alpha-1)(1-x)$, so that

$$\|\mathcal{I}_K^{\mathrm{L}}u_n\|_{L^p(0,1)}\ge(n^\alpha-1)\|(1-x)\|_{L^p(0,1)}\ge(n^\alpha-1)\|\tfrac12\|_{L^p(0,\frac12)}\ge\tfrac14(n^\alpha-1).$$

This proves that $\|\mathcal{I}_K^{\mathrm{L}} u_n\|_{L^p(0,1)} \geq (n^\alpha - 1)\gamma^{-1}\|u_n\|_{L^p(0,1)}$ with $\gamma := 1 + \frac{1}{(1-p\alpha)^{\frac{1}{p}}}$, thereby proving that $\|\mathcal{I}_K^{\mathrm{L}}\|_{\mathcal{L}(L^p;L^p)} \geq \frac{1}{4}(n^\alpha - 1)\gamma^{-1}$ for all $n \in \mathbb{N}\backslash\{0\}$. In conclusion, $\|\mathcal{I}_K^{\mathrm{L}}\|_{\mathcal{L}(L^p;L^p)} = \infty$, i.e., $\mathcal{I}_K^{\mathrm{L}}$ is not L^p stable for all $p < \frac{1}{\alpha}$.
(iii) Since α is arbitrary in $(0,1)$, the above result implies that $\mathcal{I}_K^{\mathrm{L}}$ is not L^r stable for all $r \in [1,\infty)$ in dimension one.

Exercise 11.7 (Norm scaling, $s \notin \mathbb{N}$). Let $\mathcal{A}^{\mathrm{H}}_{\lfloor m \rfloor, d} = \{\alpha \in \mathbb{N}^d \mid |\alpha| = m\}$ be the set of the multi-indices of length equal to m. We have

$$|\psi_K(v)|_{W^{s,p}(\widehat{K})} := \left(\sum_{\alpha \in \mathcal{A}^{\mathrm{H}}_{m,d}} |\partial^\alpha \psi_K(v)|^p_{W^{\sigma,p}(\widehat{K})} \right)^{\frac{1}{p}}.$$

By proceeding as in the proof of Lemma 11.7 and using that $\boldsymbol{T}_K$ is affine, we infer that the following holds true for all $\alpha \in \mathcal{A}^{\mathrm{H}}_{m,d}$:

$$\|\partial^\alpha \psi_K(v)(\widehat{\boldsymbol{x}}) - \partial^\alpha \psi_K(v)(\widehat{\boldsymbol{y}})\|_{\ell^2} \leq c\|\mathbb{A}_K\|_{\ell^2}\|\mathbb{J}_K\|^m_{\ell^2} \sum_{\beta \in \mathcal{A}^{\mathrm{H}}_{m,d}} \|(\partial^\beta v)(\boldsymbol{T}_K(\widehat{\boldsymbol{x}})) - (\partial^\beta v)(\boldsymbol{T}_K(\widehat{\boldsymbol{y}}))\|_{\ell^2}.$$

We infer that

$$\begin{aligned}
|\partial^\alpha \psi_K(v)|^p_{W^{\sigma,p}(\widehat{K})} &= \int_{\widehat{K}} \int_{\widehat{K}} \frac{\|\partial^\alpha \psi_K(v)(\widehat{\boldsymbol{x}}) - \partial^\alpha \psi_K(v)(\widehat{\boldsymbol{y}})\|^p_{\ell^2}}{\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}\|^{\sigma p + d}_{\ell^2}} \,\mathrm{d}\widehat{x}\,\mathrm{d}\widehat{y} \\
&\leq c\|\mathbb{A}_K\|^p_{\ell^2}\|\mathbb{J}_K\|^{pm}_{\ell^2} \sum_{\beta \in \mathcal{A}^{\mathrm{H}}_{m,d}} \int_{\widehat{K}} \int_{\widehat{K}} \frac{\|(\partial^\beta v)(\boldsymbol{T}_K(\widehat{\boldsymbol{x}})) - (\partial^\beta v)(\boldsymbol{T}_K(\widehat{\boldsymbol{y}}))\|^p_{\ell^2}}{\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}\|^{\sigma p + d}_{\ell^2}} \,\mathrm{d}\widehat{x}\,\mathrm{d}\widehat{y} \\
&\leq c\|\mathbb{A}_K\|^p_{\ell^2}\|\mathbb{J}_K\|^{pm}_{\ell^2} |\det(\mathbb{J}_K)|^{-2} \|\mathbb{J}_K\|^{\sigma p + d}_{\ell^2} \\
&\quad \times \sum_{\beta \in \mathcal{A}^{\mathrm{H}}_{m,d}} \int_K \int_K \frac{\|\partial^\beta v(\boldsymbol{x}) - \partial^\beta v(\boldsymbol{y})\|^p_{\ell^2}}{\|\boldsymbol{x} - \boldsymbol{y}\|^{\sigma p + d}_{\ell^2}} \,\mathrm{d}x\,\mathrm{d}y,
\end{aligned}$$

since $\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}\|_{\ell^2} = \|\mathbb{J}_K^{-1}(\boldsymbol{x} - \boldsymbol{y})\|_{\ell^2} \geq \|\mathbb{J}_K\|^{-1}_{\ell^2}\|\boldsymbol{x} - \boldsymbol{y}\|_{\ell^2}$. In conclusion, we have

$$\begin{aligned}
|\psi_K(v)|_{W^{s,p}(\widehat{K})} &\leq c\,\|\mathbb{A}_K\|_{\ell^2}\|\mathbb{J}_K\|^m_{\ell^2} |\det(\mathbb{J}_K)|^{-\frac{2}{p}} \|\mathbb{J}_K\|^{\sigma + \frac{d}{p}}_{\ell^2} |v|_{W^{s,p}(K)} \\
&\leq c\,\|\mathbb{A}_K\|_{\ell^2}\|\mathbb{J}_K\|^s_{\ell^2} |\det(\mathbb{J}_K)|^{-\frac{1}{p}} \left(|\det(\mathbb{J}_K)|^{-1}\|\mathbb{J}_K\|^d_{\ell^2}\right)^{\frac{1}{p}} |v|_{W^{s,p}(K)},
\end{aligned}$$

which proves the statement. The proof of the other inequality is similar.

Exercise 11.8 (Morrey's polynomial). We have proved in Lemma 11.9 that the map $\Phi_{k,d} : \mathbb{P}_{k,d} \to \mathbb{R}^{N_{k,d}}$ such that $\Phi_{k,d}(q) = (\int_U \partial^\alpha q \,\mathrm{d}x)_{\alpha \in \mathcal{A}_{k,d}}$ is an isomorphism. Hence, there is a unique $q \in \mathbb{P}_{k,d}$ such that $\Phi_{k,d}(q) = (\int_U \partial^\alpha u \,\mathrm{d}x)_{\alpha \in \mathcal{A}_{k,d}}$, i.e., $\int_U \partial^\alpha (u - q)\,\mathrm{d}x = 0$ for all $\alpha \in \mathcal{A}_{k,d}$. Note that the polynomial in question is denoted by $\pi(u)$ in the proof of Lemma 11.9.

Exercise 11.9 (Fractional Sobolev norm). Since $\widehat{v}$ has zero mean value over $\widehat{K}$, Lemma 3.26 implies that

$$\|\widehat{v}\|^2_{H^r(\widehat{K})} = \|\widehat{v}\|^2_{L^2(\widehat{K})} + |\widehat{v}|^2_{H^r(\widehat{K})} \leq \widehat{c}\,|\widehat{v}|^2_{H^r(\widehat{K})}.$$

Moreover, we have

$$\begin{aligned}
|\widehat{v}|^2_{H^r(\widehat{K})} &= \int_{\widehat{K}}\int_{\widehat{K}} \frac{|\widehat{v}(\widehat{\boldsymbol{x}}) - \widehat{v}(\widehat{\boldsymbol{y}})|^2}{\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}\|^{d+2r}_{\ell^2}}\,\mathrm{d}\widehat{x}\,\mathrm{d}\widehat{y} \\
&= \frac{|\widehat{K}|^2}{|K|^2}\int_K\int_K \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^2}{\|\mathbb{J}_K^{-1}(\boldsymbol{x} - \boldsymbol{y})\|^{d+2r}_{\ell^2}}\,\mathrm{d}x\,\mathrm{d}y \\
&\leq \frac{|\widehat{K}|^2}{|K|^2}\mathbb{J}_K^{d+2r}|v|^2_{H^r(K)}.
\end{aligned}$$

Hence, $\|\widehat{v}\|_{H^r(\widehat{K})} \leq c h_K^{r-\frac{d}{2}}|v|_{H^r(K)}$, where c depends on the regularity of the mesh sequence.

Chapter 12

Local inverse and functional inequalities

Exercises

Exercise 12.1 (ℓ^p vs. ℓ^r). Let p, r be two nonnegative real numbers. Let $\{a_i\}_{i\in I}$ be a finite sequence of nonnegative numbers. Set $\|a\|_{\ell^p(\mathbb{R}^I)} := (\sum_{i\in I} a_i^p)^{\frac{1}{p}}$ and $\|a\|_{\ell^r(\mathbb{R}^I)} := (\sum_{i\in I} a_i^r)^{\frac{1}{r}}$. (i) Prove that $\|a\|_{\ell^p(\mathbb{R}^I)} \le \|a\|_{\ell^r(\mathbb{R}^I)}$ for $r \le p$. (*Hint*: set $\theta_i := a_i^r/\|a\|_{\ell^r(\mathbb{R}^I)}^r$.) (ii) Prove that $\|a\|_{\ell^p(\mathbb{R}^I)} \le \mathrm{card}(I)^{\frac{r-p}{pr}} \|a\|_{\ell^r(\mathbb{R}^I)}$ for $r > p$.

Exercise 12.2 (L^p-norm of shape functions). Let $\theta_{K,i}$, $i \in \mathcal{N}$, be a local shape function. Let $p \in [1,\infty]$. Assume that $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is shape-regular. Prove that $\|\theta_{K,i}\|_{L^p(K)}$ is equivalent to $h_K^{d/p}$ uniformly w.r.t. $K \in \mathcal{T}_h$ and $h \in \mathcal{H}$.

Exercise 12.3 (dof norm). Prove Proposition 12.5. (*Hint*: use Lemma 11.7.)

Exercise 12.4 (Inverse inequality). (i) Let $k \ge 1$, $p \in [1,\infty]$, let $\widehat{K} := \{(\widehat{x}_1,\dots,\widehat{x}_d) \in (0,1)^d \mid \sum_{i\in\{1:d\}} \widehat{x}_i \le 1\}$, and set $\widehat{c}_{k,p} := \sup_{\widehat{v}\in\mathbb{P}_{k,d}} \frac{\|\nabla\widehat{v}\|_{\boldsymbol{L}^p(\widehat{K})}}{\|\widehat{v}\|_{L^p(\widehat{K})}}$. Explain why $\widehat{c}_{k,p}$ is finite. (ii) Let K be a simplex in $\mathbb{R}^d$ and let ρ_K denote the diameter of its largest inscribed ball. Show that $\|\nabla v\|_{\boldsymbol{L}^p(K)} \le \widehat{c}_{k,p}\frac{\sqrt{2}}{\rho_K}\|v\|_{L^p(K)}$ for all $v \in \mathbb{P}_{k,d}\circ\boldsymbol{T}_K$, where $\boldsymbol{T}_K : \widehat{K} \to K$ is the geometric mapping. (*Hint*: use (9.8a) and Lemma 11.1.)

Exercise 12.5 (Markov inequality). (i) Justify that the constant $C_{2,k}$ in the Markov inequality (12.7) can be determined as the largest eigenvalue of the stiffness matrix $\mathcal{A}$. (ii) Compute numerically the constant $C_{2,k}$ for $k \in \{1,2,3\}$.

Exercise 12.6 (Fractional trace inequality). Prove (12.17). (*Hint*: use a trace inequality in $W^{s,p}(\widehat{K})$.)

Exercise 12.7 (Mapped polynomial approximation). Let $(\widehat{K},\widehat{P},\widehat{\Sigma})$ be a reference finite element such $\mathbb{P}_{k,d} \subset \widehat{P}$, $k \in \mathbb{N}$. Let $\mathcal{T}_h$ be a member of a shape-regular mesh sequence. Let $\boldsymbol{T}_K(\widehat{K}) = K \in \mathcal{T}_h$ and let (K, P_K, Σ_K) be the finite element generated by the geometric mapping

$\boldsymbol{T}_K$ and the functional transformation $\psi_K(v) := \mathbb{A}_K(v \circ \boldsymbol{T}_K)$. Recall that $P_K = \psi_K^{-1}(\widehat{P})$. Show that there is c s.t.

$$\inf_{q \in P_K} |v - q|_{W^{m,p}(K)} \le c\, h_K^{r-m} |v|_{W^{r,p}(K)}, \tag{12.1}$$

for all $r \in [0, k+1]$, all $p \in [1, \infty)$ if $r \notin \mathbb{N}$ or all $p \in [1, \infty]$ if $r \in \mathbb{N}$, every integer $m \in \{0{:}\lfloor r \rfloor\}$, all $v \in W^{r,p}(K)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$, where the mesh cells are supposed to be convex sets if $r \ge 1$. (*Hint*: use Lemma 11.7 and Corollary 12.13.)

Exercise 12.8 (Trace inequality). Let U be a Lipschitz domain in $\mathbb{R}^d$. Prove that there are $c_1(U)$ and $c_2(U)$ such that $\|v\|_{L^p(\partial U)} \le c_1(U)\|v\|_{L^p(U)} + c_2(U)\|\nabla v\|_{\boldsymbol{L}^p(U)}^{\frac{1}{p}} \|v\|_{L^p(U)}^{1-\frac{1}{p}})$ for all $p \in [1, \infty)$ and all $v \in W^{1,p}(U)$. (*Hint*: accept as a fact that there exists a smooth vector field $\boldsymbol{N} \in \boldsymbol{C}^1(\overline{U})$ and $c_0(U) > 0$ such that $(\boldsymbol{N}\cdot\boldsymbol{n})_{|\partial U} \ge c_0(U)$ and $\|\boldsymbol{N}(\boldsymbol{x})\|_{\ell^2(\mathbb{R}^d)} = 1$ for all $\boldsymbol{x} \in U$.)

Exercise 12.9 (Weighted inverse inequalities). Let $k \in \mathbb{N}$. (i) Prove that $\|(1-t^2)^{\frac{1}{2}} v'\|_{L^2(-1,1)} \le \big(k(k+1)\big)^{\frac{1}{2}} \|v\|_{L^2(-1,1)}$ for all $v \in \mathbb{P}_{k,1}$. (*Hint*: let $\tilde{L}_m := \big(\frac{2m+1}{2}\big)^{1/2} L_m$, L_m being the Legendre polynomial from Definition 6.1, and prove that $\int_{-1}^1 (1-t^2)(\tilde{L}_m)'(t)(\tilde{L}_n)'(t)\,\mathrm{d}t = \delta_{mn} m(m+1)$ for every integers $m, n \in \{0{:}k\}$.) (ii) Prove that $\|v\|_{L^2(-1,1)} \le (k+2)\|(1-t^2)^{\frac{1}{2}} v\|_{L^2(-1,1)}$ for all $v \in \mathbb{P}_{k,1}$. (*Hint*: consider a Gauss–Legendre quadrature with $l_{\mathcal{Q}} := k+2$ and use the fact that the rightmost Gauss–Legendre node satisfies $\xi_{l_{\mathcal{Q}}} \le \cos(\frac{\pi}{2l_{\mathcal{Q}}})$.) *Note*: see also Verfürth [44].

Solution to exercises

Exercise 12.1 (ℓ^p vs. ℓ^r). (i) We observe that $\theta_i := a_i^r / \|a\|_{\ell^r(\mathbb{R}^I)}^r \in [0,1]$ and $\sum_{i\in I} \theta_i = 1$. Since $\frac{p}{r} \ge 1$, we infer that $\sum_{i\in I} \theta_i^{\frac{p}{r}} \le \sum_{i\in I} \theta_i = 1$. Rearranging the terms leads to the expected estimate.
(ii) Using Hölder's inequality, we infer that

$$\sum_{i\in I} \theta_i^{\frac{p}{r}} \le \left(\sum_{i\in I} \theta_i^{\frac{p}{r}\frac{r}{p}}\right)^{\frac{p}{r}} \left(\sum_{i\in I} 1^{\frac{r}{r-p}}\right)^{1-\frac{p}{r}} \le \mathrm{card}(I)^{1-\frac{p}{r}}.$$

Exercise 12.2 (L^p-norm of shape functions). Observe that

$$\|\theta_{K,i}\|_{L^p(K)} = \left(\frac{|K|}{|\widehat{K}|}\right)^{\frac{1}{p}} \|\widehat{\theta}_i\|_{L^p(\widehat{K})},$$

and use the regularity of the mesh sequence to conclude.

Exercise 12.3 (dof norm). Owing to (12.3), it is sufficient to prove the equivalence for $p = \infty$. Let $v_h = \sum_{i\in\mathcal{N}} \sigma_{K,i}(v_h)\theta_{K,i} \in P_K$. Recalling that $\theta_{K,i} = \psi_K^{-1}(\widehat{\theta}_i)$ for all $i \in \mathcal{N}$, we infer that

$$\begin{aligned}
\|v_h\|_{L^\infty(K;\mathbb{R}^q)} &\le \sum_{i\in\mathcal{N}} |\sigma_{K,i}(v_h)| \|\theta_{K,i}\|_{L^\infty(K;\mathbb{R}^q)} \\
&\le \sum_{i\in\mathcal{N}} |\sigma_{K,i}(v_h)| \|\psi_K^{-1}\|_{\mathcal{L}(L^\infty(\widehat{K};\mathbb{R}^q);L^\infty(K;\mathbb{R}^q))} \|\widehat{\theta}_i\|_{L^\infty(\widehat{K};\mathbb{R}^q)} \\
&\le c_1 \|\psi_K^{-1}\|_{\mathcal{L}(L^\infty(\widehat{K};\mathbb{R}^q);L^\infty(K;\mathbb{R}^q))} \sum_{i\in\mathcal{N}} |\sigma_{K,i}(v_h)|,
\end{aligned}$$

where $c_1 := \max_{i\in\mathcal{N}} \|\widehat{\theta}_i\|_{L^\infty(\widehat{K};\mathbb{R}^q)}$ only depends on the reference element. Using (11.7b) with $l := 0$ and $p := \infty$, we infer that

$$\|v_h\|_{L^\infty(K;\mathbb{R}^q)} \le c\,\|\mathbb{A}_K^{-1}\|_{\ell^2} \sum_{i\in\mathcal{N}} |\sigma_{K,i}(v_h)|.$$

Let us now prove the reverse bound. Let $\widehat{v}_h := \psi_K(v_h)$. Since $(\widehat{K},\widehat{P},\widehat{\Sigma})$ is a finite element, $\sum_{i\in\mathcal{N}} |\sigma_n(\widehat{v}_h)|$ is a norm on $\widehat{P}$. The equivalence of norms in $\widehat{P}$ implies that there is c_2, depending only on $(\widehat{K},\widehat{P},\widehat{\Sigma})$, such that

$$\begin{aligned}\sum_{i\in\mathcal{N}} |\sigma_{K,i}(v_h)| = \sum_{i\in\mathcal{N}} |\widehat{\sigma}_i(\widehat{v}_h)| &\le c_2\|\widehat{v}_h\|_{L^\infty(\widehat{K};\mathbb{R}^q)} = c_2\|\psi_K(v_h)\|_{L^\infty(\widehat{K};\mathbb{R}^q)}\\ &\le c_2\|\psi_K\|_{\mathcal{L}(L^\infty(K;\mathbb{R}^q);L^\infty(\widehat{K};\mathbb{R}^q))}\|v_h\|_{L^\infty(K;\mathbb{R}^q)}\\ &\le c_2'\|\mathbb{A}_K\|_{\ell^2}\|v_h\|_{L^\infty(K;\mathbb{R}^q)},\end{aligned}$$

where the last bound follows from (11.7a) with $l := 0$ and $p := \infty$. The conclusion follows from the fact that $\|\mathbb{A}_K\|_{\ell^2}\|\mathbb{A}_K^{-1}\|_{\ell^2}$ is bounded by a constant that only depends on the regularity of the mesh sequence owing to (11.12).

Exercise 12.4 (Inverse inequality). (i) Since $\mathbb{P}_{k,d}$ is finite-dimensional, the unit sphere $\widehat{S}_p := \{\widehat{v}\in\mathbb{P}_{k,d} \mid \|\widehat{v}\|_{L^p(\widehat{K})} = 1\}$ is compact. Hence, the continuous function $\widehat{v}\mapsto\|\nabla\widehat{v}\|_{\boldsymbol{L}^p(\widehat{K})}$ attains its maximum on $\widehat{S}_p$. Since the maximum in question is $\widehat{c}_{k,p}$ by definition, this proves that this real number is finite.
(ii) Let $v\in\mathbb{P}_{k,d}\circ\boldsymbol{T}_K$. Then $\widehat{v} := v\circ\boldsymbol{T}_K^{-1}\in\mathbb{P}_{k,d}$. Let $\mathbb{J}_K$ be the Jacobian of the geometric mapping $\boldsymbol{T}_K$. The chain rule (9.8a) implies that $\nabla v = \mathbb{J}_K^{-\mathsf{T}}\nabla\widehat{v}$. Since $\boldsymbol{T}_K$ is affine, we infer that

$$\begin{aligned}\|\nabla v\|_{\boldsymbol{L}^p(K)} &\le \|\mathbb{J}_K^{-1}\|_{\ell^2}|\det(\mathbb{J}_K)|^{\frac{1}{p}}\|\nabla\widehat{v}\|_{\boldsymbol{L}^p(\widehat{K})}\\ &\le \widehat{c}_{k,p}\|\mathbb{J}_K^{-1}\|_{\ell^2}|\det(\mathbb{J}_K)|^{\frac{1}{p}}\|\widehat{v}\|_{L^p(\widehat{K})}\\ &\le \widehat{c}_{k,p}\|\mathbb{J}_K^{-1}\|_{\ell^2}\|v\|_{L^p(K)}.\end{aligned}$$

Finally, invoking Lemma 11.1 gives $\|\mathbb{J}_K^{-1}\|_{\ell^2}\le\frac{h_{\widehat{K}}}{\rho_K}$. Since $\widehat{K}$ is the unit simplex and $\widehat{K}$ is convex, and we have $h_{\widehat{K}} = \sqrt{2}$ in every space dimension.

Exercise 12.5 (Markov inequality). (i) Let $v\in\mathbb{P}_{k,1}$. We can write $v(t) = \sum_{l\in\{0:k\}} v_l\tilde{L}_l(t)$. Exploiting the L^2-orthonormality of the basis and the definition of the stiffness matrix $\mathcal{A}$, we infer that

$$\frac{\|v'\|_{L^2(-1,1)}}{\|v\|_{L^2(-1,1)}} = \frac{V^{\mathsf{T}}\mathcal{A}V}{V^{\mathsf{T}}V}\le\rho(\mathcal{A}).$$

(ii) A direct computation of $\mathcal{A}$ for $k\in\{1,2,3\}$, respectively, yields

$$\begin{pmatrix}0&0\\0&3\end{pmatrix},\qquad\begin{pmatrix}0&0&0\\0&3&0\\0&0&15\end{pmatrix},\qquad\begin{pmatrix}0&0&0&0\\0&3&0&\sqrt{21}\\0&0&15&0\\0&\sqrt{21}&0&42\end{pmatrix},$$

with spectral radius 3, 15, and $\frac{45+\sqrt{1605}}{2}$, respectively.

Exercise 12.6 (Fractional trace inequality). Let $v \in W^{s,p}(K)$. Let $K \in \mathcal{T}_h$ be a mesh cell and F be a face of K. Since the mapping $\boldsymbol{T}_K$ is affine, using the trace theorem (Theorem 3.10) in $W^{s,p}(\widehat{K})$, we infer that

$$\|v\|_{L^p(F)} = \frac{|F|^{\frac1p}}{|\widehat{F}|^{\frac1p}}\|\psi_K^{\mathrm{g}}(v)\|_{L^p(\widehat{F})} \le c_{s,p}|F|^{\frac1p}(\|\psi_K^{\mathrm{g}}(v)\|_{L^p(\widehat{K})} + |\psi_K^{\mathrm{g}}(v)|_{W^{s,p}(\widehat{K})}),$$

where $c_{s,p}$ can grow unboundedly as $sp \downarrow 1$ if $p > 1$. Using Lemma 11.7, this inequality is rewritten as

$$\|v\|_{L^p(F)} \le c'_{s,p}|F|^{\frac1p}|K|^{-\frac1p}(\|v\|_{L^p(K)} + \|\mathbb{J}_K\|_{\ell^2}^{-s}|v|_{W^{s,p}(K)}).$$

The conclusion follows from the regularity of the mesh sequence (see (11.3)).

Exercise 12.7 (Mapped polynomial approximation). Let $k \in \mathbb{N}$. Let $r \in [0, k+1]$, let $p \in [1,\infty)$ if $r \notin \mathbb{N}$ or $p \in [1,\infty]$ if $r \in \mathbb{N}$, and let $m \in \{0{:}\lfloor r\rfloor\}$. Let $v \in W^{r,p}(K)$ and set $\widehat{v} := \psi_K(v)$. Let $\widehat{q}^* \in \widehat{P}$ be s.t. $|\widehat{v} - \widehat{q}^*|_{W^{m,p}(\widehat{K})} = \inf_{\widehat{q}\in\widehat{P}}|\widehat{v} - \widehat{q}|_{W^{m,p}(\widehat{K})}$. We have (the value of c changes at each occurrence)

$$\begin{aligned}
\inf_{q\in P_K}|v-q|_{W^{m,p}(K)} &\le |v - \psi_K^{-1}(\widehat{q}^*)|_{W^{m,p}(K)} = |\psi_K^{-1}(\widehat{v}) - \psi_K^{-1}(\widehat{q}^*)|_{W^{m,p}(K)}\\
&\le c\,\|\mathbb{A}_K^{-1}\|_{\ell^2}\|\mathbb{J}_K^{-1}\|_{\ell^2}^m\,|\det(\mathbb{J}_K)|^{\frac1p}\,|\widehat{v} - \widehat{q}^*|_{W^{m,p}(\widehat{K})}\\
&\le c\,\|\mathbb{A}_K^{-1}\|_{\ell^2}\|\mathbb{J}_K^{-1}\|_{\ell^2}^m\,|\det(\mathbb{J}_K)|^{\frac1p}\inf_{\widehat{q}\in\mathbb{P}_{k,d}}|\widehat{v} - \widehat{q}|_{W^{m,p}(\widehat{K})}\\
&\le c\,\|\mathbb{A}_K^{-1}\|_{\ell^2}\|\mathbb{J}_K^{-1}\|_{\ell^2}^m\,|\det(\mathbb{J}_K)|^{\frac1p}\,|\widehat{v}|_{W^{r,p}(\widehat{K})}\\
&\le c\,\|\mathbb{A}_K\|_{\ell^2}\|\mathbb{A}_K^{-1}\|_{\ell^2}\|\mathbb{J}_K\|_{\ell^2}^r\|\mathbb{J}_K^{-1}\|_{\ell^2}^m\,|v|_{W^{r,p}(K)}\\
&\le c\,h_K^{r-m}\,|v|_{W^{r,p}(K)},
\end{aligned}$$

where we used that $\psi_K^{-1}(\widehat{q}^*) \in P_K$ in the first line, (11.7b) in the second line, the definition of $\widehat{q}^*$ and $\mathbb{P}_{k,d} \subset \widehat{P}$ in the third line, Corollary 12.13 in the fourth line, (11.7a) in the fifth line, and the regularity of the mesh sequence in the last line. This proves (12.1).

Exercise 12.8 (Trace inequality). We first observe that

$$\begin{aligned}
c_0(U)\int_{\partial U}|v|^p\,\mathrm{d}x &\le \int_{\partial_U}(\boldsymbol{n}\cdot\boldsymbol{N})|v|^p\,\mathrm{d}x = \int_U \nabla\cdot(\boldsymbol{N}|v|^p)\,\mathrm{d}x\\
&\le \int_U((\nabla\cdot\boldsymbol{N})|v|^p + p(\boldsymbol{N}\cdot\nabla v)|v|^{p-1})\,\mathrm{d}x\\
&\le c_1(U)\|v\|_{L^p(U)}^p + p\|\nabla v\|_{\boldsymbol{L}^p(U)}\|v\|_{L^p(U)}^{p-1},
\end{aligned}$$

where we set $c_1(U) := \|\nabla\cdot\boldsymbol{N}\|_{L^\infty(U)}$, used that $\|\boldsymbol{N}(\boldsymbol{x})\|_{\ell^2(\mathbb{R}^d)} = 1$ for all $\boldsymbol{x} \in U$, and used Hölder's inequality to bound $\int_U \|\nabla u\|_{\ell^2}|v|^{p-1}\,\mathrm{d}x$. The conclusion follows by applying the inequality $(a+b)^{\frac1p} \le a^{\frac1p} + b^{\frac1p}$ for all $a, b \ge 0$, i.e.,

$$\|v\|_{L^p(U)} \le \left(\frac{c_1(U)}{c_0(U)}\right)^{\frac1p}\|v\|_{L^p(U)} + p^{\frac1p}c_0(U)^{-\frac1p}\|\nabla v\|_{\boldsymbol{L}^p(U)}^{\frac1p}\|v\|_{L^p(U)}^{1-\frac1p}.$$

Exercise 12.9 (Weighted inverse inequalities). (i) Without loss of generality, assume $n \le m$. Integrating by parts and since $(1-t^2)$ vanishes at $t = \pm1$, we infer that

$$\int_{-1}^1(1-t^2)(\tilde{L}_m)'(t)(\tilde{L}_n)'(t)\,\mathrm{d}t = -\int_{-1}^1\tilde{L}_m(t)\big((1-t^2)(\tilde{L}_n)'(t)\big)'\,\mathrm{d}t.$$

Since $\big((1-t^2)(\tilde{L}_n)'(t)\big)'$ is a polynomial of degree n whose leading coefficient is equal to that of $\tilde{L}_n$ multiplied by $-n(n+1)$, the orthonormality of the (normalized) Legendre polynomials implies that $\int_{-1}^1 (1-t^2)(\tilde{L}_m)'(t)(\tilde{L}_n)'(t)\,\mathrm{d}t = \delta_{mn} m(m+1)$. As a result, writing any $v \in \mathbb{P}_{k,1}$ as $v(t) = \sum_{l\in\{0:k\}} v_l \tilde{L}_l(t)$, we infer that

$$\begin{aligned}\int_{-1}^1 (1-t^2)|v'(t)|^2\,\mathrm{d}t &= \sum_{l\in\{0:k\}} v_l^2 l(l+1)\\ &\le k(k+1)\sum_{l\in\{0:k\}} v_l^2 = k(k+1)\|v\|^2_{L^2(-1,1)}.\end{aligned}$$

(ii) Since $(1-t^2)v^2$ is of degree $(2k+2)$ and the quadrature is of order $2l_{\mathcal{Q}}-1 = 2k+3$, we infer that

$$\begin{aligned}\int_{-1}^1 (1-t^2)v(t)^2\,\mathrm{d}t &= \sum_{l\in\{1:l_{\mathcal{Q}}\}} \omega_l (1-\xi_l^2) v(\xi_l)^2\\ &\ge (1-\xi_{l_{\mathcal{Q}}}^2)\sum_{l\in\{1:l_{\mathcal{Q}}\}} \omega_l v(\xi_l)^2 = (1-\xi_{l_{\mathcal{Q}}}^2)\int_{-1}^1 v(t)^2\,\mathrm{d}t.\end{aligned}$$

The conclusion follows from

$$\frac{1}{1-\xi_{l_{\mathcal{Q}}}^2} \le \frac{1}{\sin^2(\frac{\pi}{2l_{\mathcal{Q}}})} \le (l_{\mathcal{Q}})^2,$$

since $\sin(x) \ge \frac{2}{\pi}x$ for all $x \in [0, \frac{\pi}{2}]$.

Chapter 13

Local interpolation on nonaffine meshes

Exercises

Exercise 13.1 (Chain rule). Let $f \in \mathcal{C}^3(U; W_1)$ and $g \in \mathcal{C}^3(W_1; W_2)$, where V, W_1, W_2 are Banach spaces and U is an open set in V. (i) Evaluate the pure derivatives $D^2(g \circ f)(x)(h,h)$ and $D^3(g \circ f)(x)(h,h,h)$ for $x \in U$ and $h \in V$. (ii) Rewrite these expressions when f and g map from $\mathbb{R}$ to $\mathbb{R}$.

Exercise 13.2 (Pure derivatives, $\mathbb{Q}_{k,d}$-polynomials). Let $\{\boldsymbol{e}_i\}_{i\in\{1:d\}}$ be the canonical Cartesian basis of $\mathbb{R}^d$. Let $k \geq 1$. Verify that $D^{k+1}q(\boldsymbol{e}_i, \ldots, \boldsymbol{e}_i) = 0$ for all $i \in \{1{:}d\}$ if and only if $q \in \mathbb{Q}_{k,d}$. (*Hint*: by induction on d.) What is instead the characterization of polynomials in $\mathbb{P}_{k,d}$ in terms of $D^{k+1}q$?

Exercise 13.3 (Lemma 13.5). Complete the proof of Lemma 13.5 by proving (13.9) for all $m \leq k+1$. (*Hint*: use induction on m and the chain rule formula (B.4) applied to $\boldsymbol{T}^{-1}(\boldsymbol{T}(\widehat{\boldsymbol{x}}))$.)

Exercise 13.4 (Tensor-product transformation). Assume the transformation $\boldsymbol{T}$ has the tensor-product form $\boldsymbol{T}(\widehat{\boldsymbol{x}}) = \sum_{j\in\{1:d\}} t_j(\widehat{\boldsymbol{x}}_j)\boldsymbol{e}_j$ for some univariate function t_j, for all $j \in \{1{:}d\}$, where $\{\boldsymbol{e}_j\}_{j\in\{1:d\}}$ is the canonical Cartesian basis of $\mathbb{R}^d$. (i) Show that (13.15) can be sharpened as $[\![w \circ \boldsymbol{T}]\!]_{W^{l,p}(\widehat{K})} \leq c\|\det(D\widetilde{\boldsymbol{T}})^{-1}\|^{\frac{1}{p}}_{L^\infty(\widehat{K})}\|D\widetilde{\boldsymbol{T}}\|^l [\![w]\!]_{W^{l,p}(K)}$. (*Hint*: recall that $[\![w]\!]_{W^{l,p}(K)}$ is a seminorm and there exists a uniform constant c so that $\ell^l_D [\![w]\!]_{W^{l,p}(K)} \leq c\|w\|_{W^{l,p}(K)}$.) (ii) What is the consequence of this new bound on the error estimate (13.21) under the assumption (13.20)?

Exercise 13.5 ($\mathbb{Q}_1$-quadrangles). Prove that $\det(D\boldsymbol{T}(\widehat{\boldsymbol{a}}_i)) = |P_i|$, where P_i is the parallelogram formed by $\boldsymbol{a}_{i-1}$, $\boldsymbol{a}_i$, $\boldsymbol{a}_{i+1}$ (with $\boldsymbol{a}_0 := \boldsymbol{a}_4$ and $\boldsymbol{a}_5 := \boldsymbol{a}_1$). (*Hint*: see §13.5.)

Exercise 13.6 (Butterfly subdivision algorithm). Consider a mesh composed of four triangles with the connectivity array such that $\texttt{j\_geo}(1,1{:}3) := (3,4,5)$, $\texttt{j\_geo}(2,1{:}3) := (0,4,5)$, $\texttt{j\_geo}(3,1{:}3) := (1,3,5)$, $\texttt{j\_geo}(4,1{:}3) := (2,3,4)$. Let $\boldsymbol{m}$ be the midpoint of the edge $(\boldsymbol{z}_3, \boldsymbol{z}_4)$. Let $\widehat{\boldsymbol{z}}_0 := (0,0)$, $\widehat{\boldsymbol{z}}_1 := (1,0)$, $\widehat{\boldsymbol{z}}_2 := (0,1)$, $\widehat{\boldsymbol{z}}_3 := (\frac{1}{2},\frac{1}{2})$, $\widehat{\boldsymbol{z}}_4 := (0,\frac{1}{2})$, $\widehat{\boldsymbol{z}}_5 := (\frac{1}{2},0)$. Consider now the curved triangle given by the $\mathbb{P}_2$ geometric mapping $\boldsymbol{T}$ that transforms $\widehat{\boldsymbol{z}}_i$ to $\boldsymbol{z}_i$ for all $i \in \{0{:}5\}$. Let $\{f_0, \ldots, f_7\} \in \mathbb{R}$. Let $\widehat{p} \in \mathbb{P}_{2,2}$ be the polynomial defined by $\widehat{p}(\widehat{\boldsymbol{z}}_i) := f_i$ for all $i \in \{0{:}5\}$. (i)

Compute $\widehat{p}(\boldsymbol{T}^{-1}(\boldsymbol{m}))$. (ii) Consider two additional points $\boldsymbol{z}_6$, $\boldsymbol{z}_7$ and two more triangles given by $\texttt{j\_geo}(5,1{:}3) := (2,3,6)$, $\texttt{j\_geo}(6,1{:}3) := (2,4,7)$. Let $\boldsymbol{T}'$ be the $\mathbb{P}_2$ geometric mapping that transforms $\widehat{\boldsymbol{z}}_i$ to $\boldsymbol{z}_i$ for all $i \in \{2{:}7\}$. Let $\widehat{p}' \in \mathbb{P}_{2,2}$ be defined by $\widehat{p}'(\widehat{\boldsymbol{z}}_i) := f_i$ for all $i \in \{2{:}7\}$. Compute $\frac{1}{2}\big(\widehat{p}(\boldsymbol{T}^{-1}(\boldsymbol{m})) + \widehat{p}'((\boldsymbol{T}')^{-1}(\boldsymbol{m}))\big)$. *Note*: the name of the algorithm comes from the shape of the generic configuration. The algorithm is used for three-dimensional computer graphics. It allows the representation of smooth surfaces via the specification of coarser piecewise linear polygonal meshes. Given an initial polygonal mesh, a smooth surface is obtained by recursively applying the butterfly subdivision algorithm to the Cartesian coordinates of the vertices; see Dyn et al. [16].

Solution to exercises

Exercise 13.1 (Chain rule). (i) We apply Lemma B.4. For the second-order derivative, the summation in l has two terms and we obtain (we omit the point x in the (Fréchet) derivatives of f)

$$D^2(g\circ f)(x)(h,h) = Dg(f(x))(D^2 f(h,h)) + D^2 g(f(x))(Df(h),Df(h)).$$

For the third-order derivative, the summation in l has three terms and we obtain

$$\begin{aligned} D^3(g\circ f)(x)(h,h,h) = {} & Dg(f(x))(D^3 f(h,h,h)) \\ & + 3D^2 g(f(x))(Df(h),D^2 f(h,h)) \\ & + D^3 g(f(x))(Df(h),Df(h),Df(h)), \end{aligned}$$

where we used Theorem B.3 for the second term on the right-hand side.
(ii) When f and g map from $\mathbb{R}$ to $\mathbb{R}$, we obtain

$$(g\circ f)''(x) = g'(f(x))(f'(x))^2 + g''(f(x)),$$

and

$$(g\circ f)'''(x) = g'(f(x))(f'(x))^3 + 3g''(f(x))f'(x)f''(x) + g'''(f(x))(f'(x))^3.$$

Exercise 13.2 (Pure derivatives, $\mathbb{Q}_{k,d}$-polynomials). A direct verification shows that any polynomial $q \in \mathbb{Q}_{k,d}$ verifies $D^{k+1}q(\boldsymbol{x})(\boldsymbol{e}_i,\dots,\boldsymbol{e}_i) = 0$ for all $i \in \{1{:}d\}$. Conversely, assume that q is such that $D^{k+1}q(\boldsymbol{x})(\boldsymbol{e}_i,\dots,\boldsymbol{e}_i) = 0$ for all $i \in \{1{:}d\}$. We proceed by induction on d. If $d = 1$, then $q \in \mathbb{Q}_{k,1}$. For $d \geq 2$, writing $\boldsymbol{x} = (\boldsymbol{x}', x_d)$ and fixing $\boldsymbol{x}'$, we infer that the $(k+1)$-th derivative of the function $x_d \mapsto q(\boldsymbol{x}', x_d)$ is zero, so that there are functions $q_0(\boldsymbol{x}'),\dots,q_k(\boldsymbol{x}')$ s.t. $q(\boldsymbol{x}) = \sum_{m\in\{0:k\}} q_m(\boldsymbol{x}')x_d^m$. Since we have for all $j < d$,

$$0 = D^{k+1}q(\boldsymbol{x})(\boldsymbol{e}_j,\dots,\boldsymbol{e}_j) = \sum_{m\in\{0:k\}} D^{k+1}q_m(\boldsymbol{x}')(\boldsymbol{e}_j,\dots,\boldsymbol{e}_j)x_d^m,$$

and the monomials $\{x_d^m\}$ are linearly independent, we infer that

$$D^{k+1}q_m(\boldsymbol{x}')(\boldsymbol{e}_j,\dots,\boldsymbol{e}_j) = 0, \qquad \forall j \in \{1{:}d{-}1\}.$$

By the induction hypothesis, we have $q_m \in \mathbb{Q}_{k,d-1}$, so that $q \in \mathbb{Q}_{k,d}$. By proceeding as above, we finally show that $q \in \mathbb{P}_{k,d}$ if and only if $D^{k+1}q = 0$, that is, $D^{k+1}q(\boldsymbol{h}_1,\dots,\boldsymbol{h}_{k+1}) = 0$ for all $\boldsymbol{h}_1,\dots,\boldsymbol{h}_{k+1} \in \mathbb{R}^d$.

Exercise 13.3 (Lemma 13.5). Let $m \in \{1{:}k\}$. We are going to use the following induction hypothesis: For all $n \le m$, there is c_{-n} only depending on $\kappa, c_1, \dots, c_n$ so that $\|D^n(\boldsymbol{T}^{-1})\| \le c_{-n}\|D(\widetilde{\boldsymbol{T}}^{-1})\|^n$. The case $m = 1$ has been proved in Lemma 13.3 with $c_{-1} := (1 - c_1)^{-1}$. The assumption has been shown to hold true for $m = 2$ in the proof of Lemma 13.5. Let us now show that it also holds true for $m + 1$. Applying the chain rule formula (B.4) to the identity $\widehat{\boldsymbol{x}} = \boldsymbol{T}^{-1}(\boldsymbol{T}(\widehat{\boldsymbol{x}}))$ and using the triangle inequality, we obtain

$$\begin{aligned}\|D^{m+1}(\boldsymbol{T}^{-1})\| &\le c(m) \sum_{l\in\{1:m\}} \|D^l(\boldsymbol{T}^{-1})\| \\ &\qquad \times \sum_{1\le r_1+\dots+r_l=m+1} \|D^{r_1}\boldsymbol{T}\|\|D(\boldsymbol{T}^{-1})\|^{r_1}\dots\|D^{r_l}\boldsymbol{T}\|\|D(\boldsymbol{T}^{-1})\|^{r_l} \\ &\le c(m)\,\|D(\boldsymbol{T}^{-1})\|^{m+1} \sum_{l\in\{1:m\}} \|D^l(\boldsymbol{T}^{-1})\| \sum_{1\le r_1+\dots+r_l=m+1} \|D^{r_1}\boldsymbol{T}\|\dots\|D^{r_l}\boldsymbol{T}\|.\end{aligned}$$

We now use that $\|D^r\boldsymbol{T}\| \le \check{c}_r\|D\widetilde{\boldsymbol{T}}\|$ with the convention $\check{c}_1 := (1 + c_1)$ and $\check{c}_r := c_r$ for $r \ge 2$ (see (13.5) and (13.8)). We also use that $\|D(\boldsymbol{T}^{-1})\| = \|(D\boldsymbol{T})^{-1}\|$ and invoke the induction assumption. We infer that

$$\begin{aligned}\|D^{m+1}(\boldsymbol{T}^{-1})\| &\le c(m)\|D(\boldsymbol{T}^{-1})\|^{m+1} \\ &\qquad \times \sum_{l\in\{1:m\}} c_{-l}\|D(\widetilde{\boldsymbol{T}}^{-1})\|^l\|D\widetilde{\boldsymbol{T}}\|^l \sum_{1\le r_1+\dots+r_l=m+1} \check{c}_{r_1}\dots\check{c}_{r_l} \\ &\le c(m)\,\|D(\boldsymbol{T}^{-1})\|^{m+1} \sum_{l\in\{1:m\}} c_{-l}\kappa^l \sum_{1\le r_1+\dots+r_l=m+1} \check{c}_{r_1}\dots\check{c}_{r_l}.\end{aligned}$$

Setting $c_{-(m+1)} := c(m)\sum_{l\in\{1:m\}} c_{-l}\kappa^l \sum_{1\le r_1+\dots+r_l=m+1} \check{c}_{r_1}\dots\check{c}_{r_l}$ proves the assertion.

Exercise 13.4 (Tensor-product transformation). (i) When $\boldsymbol{T}$ has a tensor-product form, we obtain $D^r\boldsymbol{T}(\widehat{\boldsymbol{x}})(\boldsymbol{e}_i, \dots, \boldsymbol{e}_i) = t_i^{(r)}(x_i)\boldsymbol{e}_i$ for all $i \in \{1{:}d\}$. Therefore, using the chain rule, we now infer that

$$\begin{aligned}|D^l(w \circ \boldsymbol{T})(\widehat{\boldsymbol{x}})|_{\mathbb{Q}} &\le c \sum_{m\in\{0:l\}} |(D^m w)(\boldsymbol{T}(\widehat{\boldsymbol{x}}))|_{\mathbb{Q}} \\ &\qquad \times \sum_{1\le r_1+\dots+r_m=l} |D^{r_1}\boldsymbol{T}(\widehat{\boldsymbol{x}})|_{\mathbb{Q}}\dots|D^{r_m}\boldsymbol{T}(\widehat{\boldsymbol{x}})|_{\mathbb{Q}}.\end{aligned}$$

The expected estimate readily follows.
(ii) The error estimate (13.21) under the assumption (13.20) becomes

$$|v - \mathcal{I}_K(v)|_{W^{m,p}(K)} \le c\,\lambda^{\frac{1}{p}}\kappa^m\|D\widetilde{\boldsymbol{T}}\|^{l-m}\,[\![v]\!]_{W^{l,p}(K)}.$$

Note that such an error estimate cannot hold under the assumption (13.19) (think of $k = l = 1$, $d = 2$, and $v = x_1x_2$ for which $[\![v]\!]_{W^{1,p}(K)} = 0$).

Exercise 13.5 ($\mathbb{Q}_1$-quadrangles). Consider the (Fréchet) derivative $D\boldsymbol{T}$ at $\widehat{\boldsymbol{a}}_1$ which corresponds to $\widehat{x}_1 = \widehat{x}_2 = 0$. Then $D\boldsymbol{T}(\widehat{\boldsymbol{x}}) = (\boldsymbol{a}_2 - \boldsymbol{a}_1, \boldsymbol{a}_4 - \boldsymbol{a}_1)$. Taking into account the orientation of the enumeration of vertices leads to the expected result.

Exercise 13.6 (Butterfly subdivision algorithm). (i) Let us set $\widehat{\boldsymbol{m}} := \boldsymbol{T}^{-1}(\boldsymbol{m})$. Using the following expression of the $\mathbb{P}_2$ shape functions:

$$\begin{aligned}&\widehat{\theta}_0 = \widehat{\lambda}_0(2\widehat{\lambda}_0 - 1), && \widehat{\theta}_1 = \widehat{\lambda}_1(2\widehat{\lambda}_1 - 1), && \widehat{\theta}_2 = \widehat{\lambda}_2(2\widehat{\lambda}_2 - 1), \\ &\widehat{\theta}_3 = 4\widehat{\lambda}_1\widehat{\lambda}_2, && \widehat{\theta}_4 = 4\widehat{\lambda}_2\widehat{\lambda}_0, && \widehat{\theta}_5 = 4\widehat{\lambda}_0\widehat{\lambda}_1,\end{aligned}$$

together with $\widehat{\lambda}_0(\widehat{\boldsymbol{m}}) = \frac{1}{4}$, $\widehat{\lambda}_1(\widehat{\boldsymbol{m}}) = \frac{1}{4}$, $\widehat{\lambda}_2(\widehat{\boldsymbol{m}}) = \frac{1}{2}$, we obtain

$$\widehat{p}(\widehat{\boldsymbol{m}}) = \sum_{i=0}^{5} f_i \widehat{\theta}_i(\widehat{\boldsymbol{m}}) = -\frac{1}{8}f_0 - \frac{1}{8}f_1 + \frac{1}{2}f_3 + \frac{1}{2}f_4 + \frac{1}{4}f_5.$$

(ii) Similarly, we have

$$\widehat{p}'(\widehat{\boldsymbol{m}}) = -\frac{1}{8}f_6 - \frac{1}{8}f_7 + \frac{1}{2}f_3 + \frac{1}{2}f_4 + \frac{1}{4}f_2.$$

We infer that

$$\frac{1}{2}(\widehat{p}(\widehat{\boldsymbol{m}}) + \widehat{p}'(\widehat{\boldsymbol{m}})) = -\frac{1}{16}f_0 - \frac{1}{16}f_1 + \frac{1}{8}f_2 + \frac{1}{2}f_3 + \frac{1}{2}f_4 + \frac{1}{8}f_5 - \frac{1}{16}f_6 - \frac{1}{16}f_7.$$

The generic configuration is shown in the right panel of Figure 13.1. The mesh mapped to the reference space is shown in the left panel of Figure 13.1.

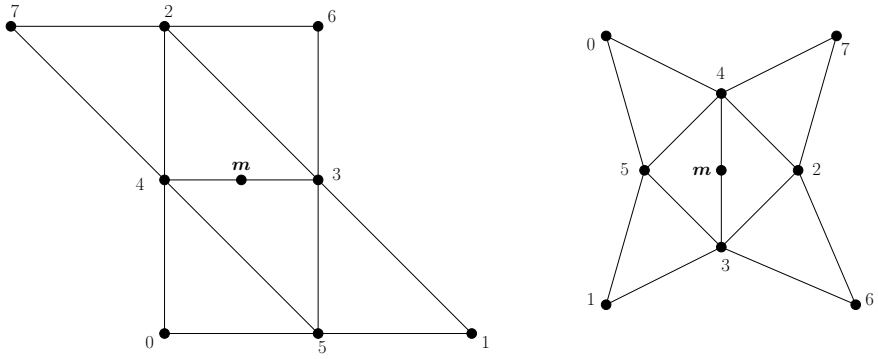


Figure 13.1: Illustration for Exercise 13.6.

Chapter 14

$\boldsymbol{H}(\mathrm{div})$ finite elements

Exercises

Exercise 14.1 ($\mathbb{RT}_{0,d}$). (i) Prove that $\int_K \iota_{F,K}\boldsymbol{\theta}_F^{\mathrm{f}}\,\mathrm{d}x = \boldsymbol{c}_F - \boldsymbol{c}_K$, where $\boldsymbol{\theta}_F^{\mathrm{f}}$ is defined in (14.3), and $\boldsymbol{c}_F$, $\boldsymbol{c}_K$ are the barycenters of F and K, respectively. (*Hint*: use (14.3) and $\int_F \boldsymbol{x}\,\mathrm{d}s = |F|\boldsymbol{c}_F$.) Provide a second proof without using (14.3). (*Hint*: fix $\boldsymbol{e}\in\mathbb{R}^d$, define $\phi(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{c}_F)\cdot\boldsymbol{e}$, observe that $\nabla\phi = \boldsymbol{e}$, and compute $\boldsymbol{e}\cdot\int_K \boldsymbol{\theta}_F^{\mathrm{f}}\,\mathrm{d}x$.) (ii) Prove that $\sum_{F\in\mathcal{F}_K}|F|\boldsymbol{\theta}_F^{\mathrm{f}}(\boldsymbol{x})\otimes\boldsymbol{n}_F = \mathbb{I}_d$ for all $\boldsymbol{x}\in K$. (*Hint*: use (7.1).) (iii) Prove that $\boldsymbol{v}(\boldsymbol{x}) = \langle\boldsymbol{v}\rangle_K + \frac{1}{d}(\nabla\cdot\boldsymbol{v})(\boldsymbol{x}-\boldsymbol{c}_K)$ for all $\boldsymbol{v}\in\mathbb{RT}_{0,d}$, where $\langle\boldsymbol{v}\rangle_K := \frac{1}{|K|}\int_K \boldsymbol{v}\,\mathrm{d}x$ is the mean value of $\boldsymbol{v}$ on K.

Exercise 14.2 ($\mathbb{RT}_{0,d}$ in 3D). Let $d=3$. Let F_i, $i\in\{0{:}3\}$, be a face of K with vertices $\{\boldsymbol{a}_r,\boldsymbol{a}_p,\boldsymbol{a}_q\}$ s.t. $\big((\boldsymbol{z}_q-\boldsymbol{z}_r)\times(\boldsymbol{z}_p-\boldsymbol{z}_r)\big)\cdot\boldsymbol{n}_{K|F_i} > 0$. (i) Prove that $\nabla\lambda_p\times\nabla\lambda_q = \frac{\boldsymbol{z}_r-\boldsymbol{z}_i}{6|K|}$ and prove similar formulas for $\nabla\lambda_q\times\nabla\lambda_r$ and $\nabla\lambda_r\times\nabla\lambda_p$. (*Hint*: prove the formula in the reference simplex, then use Exercise 9.5.) (ii) Prove that $\boldsymbol{\theta}_i^{\mathrm{f}} = -2(\lambda_p\nabla\lambda_q\times\nabla\lambda_r + \lambda_q\nabla\lambda_r\times\nabla\lambda_p + \lambda_r\nabla\lambda_p\times\nabla\lambda_q)$. Find the counterpart of this formula if $d=2$.

Exercise 14.3 (Piola transformation). (i) Let $\boldsymbol{v}\in\boldsymbol{C}^1(K)$ and $q\in C^0(K)$. Prove that $\int_K q\,\nabla\cdot\boldsymbol{v}\,\mathrm{d}x = \int_{\widehat{K}}\psi_K^{\mathrm{g}}(q)\nabla\cdot\boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{v})\,\mathrm{d}\widehat{x}$. (ii) Show that $\int_K \boldsymbol{v}\cdot\boldsymbol{\theta}\,\mathrm{d}x = \epsilon_K\int_{\widehat{K}}\boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{v})\cdot\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{\theta})\,\mathrm{d}\widehat{x}$ for all $\boldsymbol{\theta}\in\boldsymbol{C}^1(K)$.

Exercise 14.4 (Generating $\mathbb{RT}_{k,d}$). (i) Let $\boldsymbol{c}\in\mathbb{R}^d$, $q\in\mathbb{P}_{k,d}^{\mathrm{H}}$, and $\mathbb{A}\in\mathbb{R}^{d\times d'}$. Show that there is $r\in\mathbb{P}_{k-1,d'}$ such that $q(\mathbb{A}\boldsymbol{y}+\boldsymbol{c}) = q(\mathbb{A}\boldsymbol{y}) + r(\boldsymbol{y})$. (ii) Defining $s(\boldsymbol{y}) := q(\mathbb{A}\boldsymbol{y})$, show that $s\in\mathbb{P}_{k,d'}^{\mathrm{H}}$. (iii) Prove that $(\boldsymbol{\psi}_K^{\mathrm{d}})^{-1}(\mathbb{RT}_{k,d})\subset\mathbb{RT}_{k,d}$. (iv) Prove the converse inclusion.

Exercise 14.5 (BDM). Verify that $\mathrm{card}(\Sigma) = \dim(\mathbb{P}_{k,d})$ for $d\in\{2,3\}$.

Exercise 14.6 (Cartesian Raviart–Thomas element). (i) Propose a basis for $\mathbb{RT}_{0,2}^{\square}$ and for $\mathbb{RT}_{0,3}^{\square}$ in $K := [0,1]^d$. (ii) Prove (14.15). (iii) Prove Proposition 14.24.

Solution to exercises

Exercise 14.1 ($\mathbb{RT}_{0,d}$). (i) By definition, we have $\int_K \iota_{F,K}\boldsymbol{\theta}_F^{\mathrm{f}}\,\mathrm{d}x = \frac{1}{d}(\boldsymbol{c}_K - \boldsymbol{z}_F)$ since $\int_K \boldsymbol{x}\,\mathrm{d}x = |K|\boldsymbol{c}_K$. Let us prove that $\boldsymbol{c}_K - \boldsymbol{z}_F = d(\boldsymbol{c}_F - \boldsymbol{c}_K)$. Since $\boldsymbol{c}_K = \frac{1}{d+1}\sum_{F\in\mathcal{F}_K}\boldsymbol{z}_F$, we infer that

$$\begin{aligned} d(\boldsymbol{c}_F - \boldsymbol{c}_K) &= \left(\sum_{F'\in\mathcal{F}_K\setminus\{F\}} \boldsymbol{z}_{F'}\right) - d\boldsymbol{c}_K \\ &= \left(\sum_{F'\in\mathcal{F}_K} \boldsymbol{z}_{F'}\right) - \boldsymbol{z}_F - d\boldsymbol{c}_K \\ &= (d+1)\boldsymbol{c}_K - \boldsymbol{z}_F - d\boldsymbol{c}_K = \boldsymbol{c}_K - \boldsymbol{z}_F. \end{aligned}$$

Hence, we have

$$\int_K \iota_{F,K}\boldsymbol{\theta}_F^{\mathrm{f}}\,\mathrm{d}x = \frac{1}{d}(\boldsymbol{c}_K - \boldsymbol{z}_F) = \boldsymbol{c}_F - \boldsymbol{c}_K.$$

For the second proof, let $\boldsymbol{e}\in\mathbb{R}^d$. Let $\phi(\boldsymbol{x}) \coloneqq (\boldsymbol{x}-\boldsymbol{c}_F)\cdot\boldsymbol{e}$ and observe that $\nabla\phi = \boldsymbol{e}$. This gives

$$\boldsymbol{e}\cdot\int_K \boldsymbol{\theta}_F^{\mathrm{f}}\,\mathrm{d}x = \int_K \boldsymbol{\theta}_F^{\mathrm{f}}\cdot\nabla\phi\,\mathrm{d}x = -\int_K \phi\nabla\cdot\boldsymbol{\theta}_F^{\mathrm{f}}\,\mathrm{d}x + \sum_{F'\in\mathcal{F}_K}\int_{F'}(\boldsymbol{\theta}_F^{\mathrm{f}}\cdot\boldsymbol{n}_{K|F'})\phi\,\mathrm{d}s.$$

Owing to Lemma 14.7, $\boldsymbol{\theta}_F^{\mathrm{f}}\cdot\boldsymbol{n}_K$ is piecewise constant with $\boldsymbol{\theta}_F^{\mathrm{f}}\cdot\boldsymbol{n}_{K|F'} = \iota_{F,K}\frac{\delta_{FF'}}{|F|}$. Moreover, we have $|K|\nabla\cdot\boldsymbol{\theta}_F^{\mathrm{f}} = \int_K \nabla\cdot\boldsymbol{\theta}_F^{\mathrm{f}}\,\mathrm{d}x = \int_F \boldsymbol{\theta}_F^{\mathrm{f}}\cdot\boldsymbol{n}_{K|F}\,\mathrm{d}s = \iota_{F,K}$. We infer that

$$\boldsymbol{e}\cdot\int_K \iota_{F,K}\boldsymbol{\theta}_F^{\mathrm{f}}\,\mathrm{d}x = -\frac{1}{|K|}\int_K \phi\,\mathrm{d}x + \frac{1}{|F|}\int_F \phi\,\mathrm{d}s = -(\boldsymbol{c}_K - \boldsymbol{c}_F)\cdot\boldsymbol{e},$$

since $\int_K \phi\,\mathrm{d}x = \phi(\boldsymbol{c}_K)|K|$ and $\int_F \phi\,\mathrm{d}s = 0$. This implies that $\int_K \iota_{F,K}\boldsymbol{\theta}_F^{\mathrm{f}}\,\mathrm{d}x = \boldsymbol{c}_F - \boldsymbol{c}_K$ since the above equality holds true for all $\boldsymbol{e}\in\mathbb{R}^d$.
(ii) Let $\boldsymbol{x}\in K$. We observe that

$$\begin{aligned} \sum_{F\in\mathcal{F}_K}|F|\boldsymbol{\theta}_F^{\mathrm{f}}(\boldsymbol{x})\otimes\boldsymbol{n}_F &= \sum_{F\in\mathcal{F}_K}|F|\iota_{F,K}\boldsymbol{\theta}_F^{\mathrm{f}}(\boldsymbol{x})\otimes\boldsymbol{n}_{K|F} \\ &= \sum_{F\in\mathcal{F}_K}\frac{|F|}{d|K|}(\boldsymbol{x}-\boldsymbol{z}_Z)\otimes\boldsymbol{n}_{K|F} \\ &= \sum_{F\in\mathcal{F}_K}\frac{|F|}{d|K|}(\boldsymbol{c}_K-\boldsymbol{z}_i)\otimes\boldsymbol{n}_{K|F} \\ &= \sum_{i\in\mathcal{F}_K}\frac{|F|}{|K|}(\boldsymbol{c}_F-\boldsymbol{c}_K)\otimes\boldsymbol{n}_{K|F} = \mathbb{I}_d, \end{aligned}$$

where we used the definition of $\boldsymbol{\theta}_F^{\mathrm{f}}$, the first geometric identity in (7.1) to replace $\boldsymbol{x}$ by $\boldsymbol{c}_K$, the fact that $\boldsymbol{c}_K - \boldsymbol{z}_F = d(\boldsymbol{c}_F - \boldsymbol{c}_K)$, and the second geometric identity in (7.1) to conclude.
(iii) Let $\boldsymbol{v}\in\mathbb{RT}_{0,d}$. We can write $\boldsymbol{v} = \boldsymbol{a} + b(\boldsymbol{x}-\boldsymbol{c}_K)$, where $\boldsymbol{a}\in\mathbb{R}^d$, $b\in\mathbb{R}$, whence we infer that $\nabla\cdot\boldsymbol{v} = bd$, i.e., $b = \frac{1}{d}\nabla\cdot\boldsymbol{v}$. Moreover, since $(\boldsymbol{x}-\boldsymbol{c}_K)$ has zero mean value on K, we infer that $\boldsymbol{a} = \langle\boldsymbol{v}\rangle_K$. In conclusion, $\boldsymbol{v} = \langle\boldsymbol{v}\rangle_K + \frac{1}{d}(\nabla\cdot\boldsymbol{v})(\boldsymbol{x}-\boldsymbol{c}_K)$.

Exercise 14.2 ($\mathbb{RT}_{0,d}$ in 3D). (i) Let us first notice that the assumption that $\big((\boldsymbol{z}_q-\boldsymbol{z}_r)\times(\boldsymbol{z}_p-\boldsymbol{z}_r)\big)\cdot\boldsymbol{n}_{K|F_i} > 0$ means that the vectors $(\boldsymbol{z}_p-\boldsymbol{z}_r)$, $(\boldsymbol{z}_q-\boldsymbol{z}_r)$, $(\boldsymbol{z}_i-\boldsymbol{z}_r)$ form a right-handed triple.

Let us do the computation in the reference simplex. Let $\widehat{\boldsymbol{z}}_r := \boldsymbol{0}$, $\widehat{\boldsymbol{z}}_p - \widehat{\boldsymbol{z}}_r := (1,0,0)^\mathsf{T}$, $\widehat{\boldsymbol{z}}_q - \widehat{\boldsymbol{z}}_r := (0,1,0)^\mathsf{T}$, and $\widehat{\boldsymbol{z}}_i - \widehat{\boldsymbol{z}}_r := (0,0,1)^\mathsf{T}$. Then $\widehat{\lambda}_p = \widehat{x}_1$, $\widehat{\lambda}_q = \widehat{x}_2$, $\nabla\widehat{\lambda}_p = (1,0,0)^\mathsf{T}$, and $\widehat{\lambda}_q = (0,1,0)^\mathsf{T}$. This implies that $\nabla\widehat{\lambda}_p\times\nabla\widehat{\lambda}_q = (0,0,1) = \widehat{\boldsymbol{z}}_i - \widehat{\boldsymbol{z}}_r$. Since $6|\widehat{K}| = 1$, we infer that $\nabla\widehat{\lambda}_p\times\nabla\widehat{\lambda}_q = \frac{\widehat{\boldsymbol{z}}_i-\widehat{\boldsymbol{z}}_r}{6|\widehat{K}|}$. Let us now prove the formula in K. Let $\boldsymbol{T}_K$ be the affine mapping that transforms $(\widehat{\boldsymbol{z}}_p, \widehat{\boldsymbol{z}}_q, \widehat{\boldsymbol{z}}_r, \widehat{\boldsymbol{z}}_i)$ into $(\boldsymbol{z}_p, \boldsymbol{z}_q, \boldsymbol{z}_r, \boldsymbol{z}_i)$. Let $\mathbb{J}_K$ be the Jacobian matrix of $\boldsymbol{T}_K$. Observe that $\det(\mathbb{J}_K) > 0$ since $(\widehat{\boldsymbol{z}}_p - \widehat{\boldsymbol{z}}_r)$, $(\widehat{\boldsymbol{z}}_q - \widehat{\boldsymbol{z}}_r)$, $(\widehat{\boldsymbol{z}}_i - \widehat{\boldsymbol{z}}_r)$ and $(\boldsymbol{z}_p - \boldsymbol{z}_r)$, $(\boldsymbol{z}_q - \boldsymbol{z}_r)$, $(\boldsymbol{z}_i - \boldsymbol{z}_r)$ form two right-handed triples. Owing to Exercise 9.5, we infer that

$$\begin{aligned}\nabla\lambda_p\times\nabla\lambda_q &= (\mathbb{J}_K^{-\mathsf{T}}\nabla\widehat{\lambda}_p)\times(\mathbb{J}_K^{-\mathsf{T}}\nabla\widehat{\lambda}_q) = \det(\mathbb{J}_K^{-\mathsf{T}})\mathbb{J}_K(\nabla\widehat{\lambda}_p\times\nabla\widehat{\lambda}_q)\\ &= \det(\mathbb{J}_K^{-\mathsf{T}})\mathbb{J}_K\frac{\widehat{\boldsymbol{z}}_i - \widehat{\boldsymbol{z}}_r}{6|\widehat{K}|} = \det(\mathbb{J}_K)^{-1}\frac{\boldsymbol{z}_i - \boldsymbol{z}_r}{6|\widehat{K}|},\end{aligned}$$

which proves that $\nabla\lambda_p\times\nabla\lambda_q = \frac{\boldsymbol{z}_i-\boldsymbol{z}_r}{6|K|}$ since $\det(\mathbb{J}_K) = \frac{|K|}{|\widehat{K}|}$. By circular permutation on the indices (p,q,r) (which does not change the orientation of K), we also have $\nabla\lambda_q\times\nabla\lambda_r = \frac{\boldsymbol{z}_i-\boldsymbol{z}_p}{6|K|}$ and $\nabla\lambda_r\times\nabla\lambda_p = \frac{\boldsymbol{z}_i-\boldsymbol{z}_q}{6|K|}$.
(ii) Recall that $\boldsymbol{\theta}_i^{\mathrm{f}} = \frac{\boldsymbol{x}-\boldsymbol{z}_i}{3|K|}$ and that

$$\boldsymbol{x} - \boldsymbol{z}_i = \lambda_p(\boldsymbol{x})(\boldsymbol{z}_p - \boldsymbol{z}_i) + \lambda_q(\boldsymbol{x})(\boldsymbol{z}_q - \boldsymbol{z}_i) + \lambda_r(\boldsymbol{x})(\boldsymbol{z}_r - \boldsymbol{z}_i).$$

It follows immediately from Step (i) that

$$\boldsymbol{\theta}_i^{\mathrm{f}}(\boldsymbol{x}) = -2(\lambda_p(\boldsymbol{x})\nabla\lambda_q\times\nabla\lambda_r + \lambda_q(\boldsymbol{x})\nabla\lambda_r\times\nabla\lambda_p + \lambda_r(\boldsymbol{x})\nabla\lambda_p\times\nabla\lambda_q).$$

Exercise 14.3 (Piola transformation). (i) This identity follows from (9.8c), i.e., $\nabla\cdot\boldsymbol{v}(x) = \frac{1}{\det(\mathbb{J}_K(\widehat{\boldsymbol{x}}))}\nabla\cdot\boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{v})(\widehat{\boldsymbol{x}})$.
(ii) We prove the second identity as follows:

$$\begin{aligned}\int_K \boldsymbol{v}\cdot\boldsymbol{\theta}\,\mathrm{d}x &= \int_{\widehat{K}}(\boldsymbol{v}\circ\boldsymbol{T}_K)\cdot(\boldsymbol{\theta}\circ\boldsymbol{T}_K)|\det(\mathbb{J}_K)|\,\mathrm{d}\widehat{x}\\ &= \epsilon_K\int_{\widehat{K}}(\det(\mathbb{J}_K)\mathbb{J}_K^{-1}\boldsymbol{v}\circ\boldsymbol{T}_K)\cdot(\mathbb{J}_K^{\mathsf{T}}\boldsymbol{\theta}\circ\boldsymbol{T}_K)\,\mathrm{d}\widehat{x}\\ &= \epsilon_K\int_{\widehat{K}}\boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{v})\cdot\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{\theta})\,\mathrm{d}\widehat{x}.\end{aligned}$$

Exercise 14.4 (Generating $\mathbb{RT}_{k,d}$). (i) Let $\boldsymbol{x}, \boldsymbol{c} \in \mathbb{R}^d$ and consider the polynomial $q(\boldsymbol{x}) := \sum_{|\alpha|=d} a_\alpha x_1^{\alpha_1}\dots x_d^{\alpha_d}$. We have

$$\begin{aligned}q(\boldsymbol{x}+\boldsymbol{c}) &= \sum_{|\alpha|=d} a_\alpha(x_1+c_1)^{\alpha_1}\dots(x_d+c_d)^{\alpha_d}\\ &= \sum_{|\alpha|=d} a_\alpha(x_1^{\alpha_1} + r_1(x_1))\dots(x_d^{\alpha_d} + r_d(x_d)),\end{aligned}$$

where $r_i \in \mathbb{P}_{\alpha_i-1,d}$ for all $i \in \{1{:}d\}$. We infer that

$$q(\boldsymbol{x}+\boldsymbol{c}) = \sum_{|\alpha|=d} a_\alpha x_1^{\alpha_1}\dots x_d^{\alpha_d} + t(\boldsymbol{x}) = q(\boldsymbol{x}) + t(\boldsymbol{x}),$$

where $t \in \mathbb{P}_{k-1,d}$. Replacing $\boldsymbol{x}$ by $\mathbb{A}\boldsymbol{y}$, we obtain

$$q(\mathbb{A}\boldsymbol{y}+\boldsymbol{c}) = q(\mathbb{A}y) + t(\mathbb{A}\boldsymbol{y}).$$

But defining r such that $r(\boldsymbol{y}) = t(\mathbb{A}\boldsymbol{y})$, we have $r \in \mathbb{P}_{k-1,d'}$.
(ii) Let

$$p_i(\boldsymbol{y}) := \left(\sum_{j \in \{1:d'\}} \mathbb{A}_{ij} \boldsymbol{y}_j \right)^{\alpha_i}.$$

This polynomial is homogeneous of degree α_i. Moreover, the product of a homogeneous polynomial of degree α_i with a homogeneous polynomial of degree α_j is a homogeneous polynomial of degree $\alpha_i + \alpha_j$. Hence, the polynomial

$$q(\mathbb{A}\boldsymbol{y}) = \sum_{|\alpha|=d} a_\alpha p_1(\boldsymbol{y}) \dots p_d(\boldsymbol{y})$$

is homogeneous of degree $\alpha_1 + \ldots + \alpha_d = |\alpha| = k$.
(iii) Let $\boldsymbol{T}_K(\widehat{\boldsymbol{x}}) := \mathbb{J}_K \widehat{\boldsymbol{x}} + \boldsymbol{b}_K$ with $\mathbb{J}_K \in \mathbb{R}^{d\times d}$ and $\boldsymbol{b}_K \in \mathbb{R}^d$. Let $\boldsymbol{v}$ be a member of $(\boldsymbol{\psi}_K^{\mathrm{d}})^{-1}(\mathbb{RT}_{k,d})$. Then, $\boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{v}) = \widehat{\boldsymbol{p}} + \widehat{\boldsymbol{x}}\widehat{q}$ with $\widehat{\boldsymbol{p}} \in \mathbb{P}_{k,d}$ and $\widehat{q} \in \mathbb{P}_{k,d}^{\mathrm{H}}$, yielding

$$\boldsymbol{v} = (\boldsymbol{\psi}_K^{\mathrm{d}})^{-1}(\widehat{\boldsymbol{p}} + \widehat{\boldsymbol{x}}\widehat{q}) = \frac{1}{\det(\mathbb{J}_K)} \mathbb{J}_K (\widehat{\boldsymbol{p}} \circ \boldsymbol{T}_K^{-1} + (\widehat{\boldsymbol{x}}\widehat{q}) \circ \boldsymbol{T}_K^{-1}).$$

Using $\widehat{\boldsymbol{x}} = \mathbb{J}_K^{-1}(\boldsymbol{x} - \boldsymbol{b}_K)$, we have

$$\widehat{q} \circ \boldsymbol{T}_K^{-1} = \widehat{q}(\mathbb{J}_K^{-1}\boldsymbol{x} - \mathbb{J}_K^{-1}\boldsymbol{b}_K) = \widehat{q}(\mathbb{J}_K^{-1}\boldsymbol{x}) + r,$$

where $r \in \mathbb{P}_{k-1,d}$, and we have shown that $\widehat{q} \circ \mathbb{J}_K^{-1} \in \mathbb{P}_{k,d}^{\mathrm{H}}$. Hence, we have

$$\boldsymbol{v} = \boldsymbol{s} + \frac{1}{\det(\mathbb{J}_K)} \mathbb{J}_K \mathbb{J}_K^{-1} \boldsymbol{x} (\widehat{q} \circ \mathbb{J}_K^{-1}) = \boldsymbol{s} + \boldsymbol{x}t,$$

where $\boldsymbol{s} \in \mathbb{P}_{k,d}$ and $t \in \mathbb{P}_{k,d}^{\mathrm{H}}$. We conclude that $(\boldsymbol{\psi}_K^{\mathrm{d}})^{-1}(\mathbb{RT}_{k,d}) \subset \mathbb{RT}_{k,d}$.
(iv) The converse inclusion follows from a dimension argument.

Exercise 14.5 (BDM). For $d = 2$, we have $\mathrm{card}(\Sigma) = 3(k+1) + (k-1)(k+1) = (k+1)(k+2) = \dim(\mathbb{P}_{k,2})$. For $d = 3$, we have $\mathrm{card}(\Sigma) = 4\frac{1}{2}(k+1)(k+2) + \frac{1}{2}(k-1)(k+1)(k+2) = \frac{1}{2}(k+1)(k+2)(k+3) = \dim(\mathbb{P}_{k,3})$.

Exercise 14.6 (Cartesian Raviart–Thomas element). (i) A basis for $\mathbb{RT}_{0,2}^{\square}$ is

$$\left\{ \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right), \left(\begin{smallmatrix}x_1\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\x_2\end{smallmatrix}\right) \right\},$$

whereas a basis for $\mathbb{RT}_{0,3}^{\square}$ is

$$\left\{ \left(\begin{smallmatrix}1\\0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\1\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right), \left(\begin{smallmatrix}x_1\\0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\x_2\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\0\\x_3\end{smallmatrix}\right) \right\}.$$

(ii) Let $v_1 \in \mathbb{Q}_{k+1,k,\ldots,k}$ so that $v_1(\boldsymbol{x}) = \sum_{\alpha \in \mathcal{A}_{1,k,d}} a_\alpha x_1^{\alpha_1} \dots x_d^{\alpha_d}$, where $\mathcal{A}_{1,k,d} := \{(\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d \mid \alpha_1 \le k+1, \alpha_2, \dots, \alpha_d \le k\}$. Hence, $\partial_1 v_1(\boldsymbol{x}) = \sum_\alpha a_\alpha \alpha_1 x_1^{\alpha_1 - 1} \dots x_d^{\alpha_d} \in \mathbb{Q}_{k,d}$. The same reasoning on the other indices implies that $\nabla\cdot(\mathbb{RT}_{k,d}^{\square}) \subset \mathbb{Q}_{k,d}$. Let us prove that $\boldsymbol{v}_{|H}\cdot\boldsymbol{n}_H \in \mathbb{Q}_{k,d-1} \circ \boldsymbol{T}_H^{-1}$ for all $\boldsymbol{v} \in \mathbb{RT}_{k,d}^{\square}$. We do the proof for $\boldsymbol{n}_H = \boldsymbol{e}_1$, which means that x_1 is constant over H. Hence, we have

$$\boldsymbol{v}_{|H}\cdot\boldsymbol{n}_H = v_{1|H} = \sum_{\alpha \in \mathcal{A}_{1,k,d}} (a_\alpha x_1^{\alpha_1}) x_2^{\alpha_2} \dots x_d^{\alpha_d} = \sum_{\beta \in \mathcal{B}_{k,d}} b_\beta x_2^{\beta_1} \dots x_d^{\beta_{d_1}},$$

where $\mathcal{B}_{k,d} \coloneqq \{(\beta_1, \dots, \beta_{d-1}) \in \mathbb{N}^{d-1} \mid \beta_1, \dots, \beta_{d-1} \le k\}$. Let $\boldsymbol{T}_H : \mathbb{R}^{d-1} \to H$ be defined by $\boldsymbol{T}_H(y_1, \dots, y_{d-1}) \coloneqq (x_1, y_1, \dots, y_{d-1})$. Then $\boldsymbol{T}_H^{-1}(\boldsymbol{x}) = (x_2, \dots, x_d)$. Let us define the function $q(\boldsymbol{y}) \coloneqq \sum_{\beta \in \mathcal{B}_{k,d}} b_\beta y_1^{\beta_1} \dots y_d^{\beta_{d_1}}$. Then $\boldsymbol{v}_{|H} \cdot \boldsymbol{n}_H = q \circ \boldsymbol{T}_H^{-1}$ where $q \in \mathbb{Q}_{k,d-1}$.
(iii) Observe first that

$$\operatorname{card}(\Sigma) = dk(k+1)^{d-1} + 2d(k+1)^{d-1} = d(k+1)^{d-1}(k+2) = \dim(\mathbb{RT}_k^{\square}).$$

Let $\boldsymbol{v} \in \mathbb{RT}_k^{\square}$ be such that $\sigma(\boldsymbol{v}) = 0$ for all $\sigma \in \Sigma$. The assumption $\sigma_{i,m}^{\mathrm{f}}(\boldsymbol{v}) = 0$, for all $i \in \{1{:}2d\}$ and all $m \in \{1{:}n_{\mathrm{sh}}^{\mathrm{f}}\}$, together with the fact that $\boldsymbol{v}_{|F_i} \cdot \boldsymbol{n}_{F_i} \in \mathbb{Q}_{k,d-1} \circ \boldsymbol{T}_{F_i}^{-1}$, implies that $\boldsymbol{v}_{|F_i} \cdot \boldsymbol{n}_{F_i} = 0$. This, in turn, implies that $\boldsymbol{v}$ can be rewritten as $\boldsymbol{v} = (x_1(1-x_1)r_1, \dots, x_d(1-x_d)r_d)^{\mathsf{T}}$, where $\boldsymbol{r} \coloneqq (r_1, \dots, r_d)^{\mathsf{T}}$ is a member of $\mathbb{Q}_{k-1,k,\dots,k} \times \dots \times \mathbb{Q}_{k,\dots,k,k-1}$. Then the assumption $\sigma_{i,m}^{\mathrm{c}}(\boldsymbol{v}) = 0$ for all $i \in \{1{:}d\}$ and all $m \in \{1{:}n_{\mathrm{sh}}^{\mathrm{c}}\}$ implies that $\int_K \boldsymbol{v} \cdot \boldsymbol{r} \, \mathrm{d}x = 0$, which, in turn, leads to $\boldsymbol{r} = 0$, thereby proving that $\boldsymbol{v} = 0$.

Chapter 15

$\boldsymbol{H}(\mathrm{curl})$ finite elements

Exercises

Exercise 15.1 ($\mathbb{S}_{1,d}$). (i) Prove that for all $\boldsymbol{q} \in \mathbb{S}_{1,d}$, there is a unique skew-symmetric matrix $\mathbb{Q}$ s.t. $\boldsymbol{q}(\boldsymbol{x}) = \mathbb{Q}\boldsymbol{x}$. (ii) Propose a basis of $\mathbb{S}_{1,d}$. (iii) Show that $\boldsymbol{q} \in \mathbb{S}_{1,3}$ if and only if there is $\boldsymbol{b} \in \mathbb{R}^3$ such that $\boldsymbol{q}(\boldsymbol{x}) = \boldsymbol{b}\times\boldsymbol{x}$.

Exercise 15.2 (Cross product). (i) Prove that $(\mathbb{A}\boldsymbol{b})\times(\mathbb{A}\boldsymbol{c}) = \mathbb{A}(\boldsymbol{b}\times\boldsymbol{c})$ for every rotation matrix $\mathbb{A} \in \mathbb{R}^{3\times3}$ and all $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^3$. (*Hint*: use Exercise 9.5.) (ii) Show that $(\boldsymbol{a}\times\boldsymbol{b})\times\boldsymbol{c} = (\boldsymbol{a}\cdot\boldsymbol{c})\boldsymbol{b} - (\boldsymbol{b}\cdot\boldsymbol{c})\boldsymbol{a}$. (*Hint*: $(\boldsymbol{a}\times\boldsymbol{b})_k = \varepsilon_{ikj}a_ib_j$ with Levi-Civita tensor ε_{ikj}; see also the proof of Lemma 9.6.) (iii) Prove that $-(\boldsymbol{b}\times\boldsymbol{n})\times\boldsymbol{n} + (\boldsymbol{b}\cdot\boldsymbol{n})\boldsymbol{n} = \boldsymbol{b}$ if $\boldsymbol{n}$ is a unit vector.

Exercise 15.3 ($\mathbb{N}_{0,3}$). (i) Prove (15.4). (*Hint*: verify that $\boldsymbol{t}_E\cdot\nabla\lambda_q = 1$ and $\boldsymbol{t}_E\cdot\nabla\lambda_p = -1$.) (ii) Prove that $\boldsymbol{v} = \langle\boldsymbol{v}\rangle_K + \frac{1}{2}(\nabla\times\boldsymbol{v})\times(\boldsymbol{x} - \boldsymbol{c}_K)$ for all $\boldsymbol{v} \in \mathbb{N}_{0,3}$, where $\langle\boldsymbol{v}\rangle_K$ is the mean value of $\boldsymbol{v}$ on K and $\boldsymbol{c}_K$ is the barycenter of K. (*Hint*: $\nabla\times(\boldsymbol{b}\times\boldsymbol{x}) = 2\boldsymbol{b}$ for $\boldsymbol{b} \in \mathbb{R}^3$.) (iii) Let $\boldsymbol{\theta}^{\mathrm{e}}_E$ be the shape function associated with the edge $E \in \mathcal{E}_K$. Let $F \in \mathcal{F}_K$ with unit normal $\boldsymbol{n}_{K|F}$ pointing outward K. Prove that $(\boldsymbol{\theta}^{\mathrm{e}}_E)_{|F}\times\boldsymbol{n}_{K|F} = \boldsymbol{0}$ if E is not an edge of F, and $\int_F \boldsymbol{\theta}^{\mathrm{e}}_E\times\boldsymbol{n}_{K|F}\,\mathrm{d}s = \iota_{E,F}(\boldsymbol{c}_E - \boldsymbol{c}_F)$ otherwise, where $\boldsymbol{c}_E$ is the barycenter of E, $\boldsymbol{c}_F$ that of F, and $\iota_{E,F} = -1$ if $\boldsymbol{n}_{K|F}\times\boldsymbol{t}_E$ points outward F, $\iota_{E,F} = 1$ otherwise. (*Hint*: use Lemma 15.15 and Exercise 14.1(ii).) (iv) Let $\mathcal{F}_E$ collect the two faces sharing $E \in \mathcal{E}_K$. Prove that $\int_K \boldsymbol{\theta}^{\mathrm{e}}_E\,\mathrm{d}x = \frac{1}{2}\sum_{F\in\mathcal{F}_E}\iota_{E,F}(\boldsymbol{c}_F - \boldsymbol{c}_K)\times(\boldsymbol{c}_E - \boldsymbol{c}_F)$. (*Hint*: take the inner product with an arbitrary vector $\boldsymbol{e} \in \mathbb{R}^3$ and introduce the function $\boldsymbol{\psi}(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{e}\times(\boldsymbol{x} - \boldsymbol{c}_K)$.)

Exercise 15.4 (Rotated $\mathbb{RT}_{k,2}$). Prove Lemma 15.9. (*Hint*: observe that $\boldsymbol{R}_{\frac{\pi}{2}}(\mathbb{P}_{k,2}) = \mathbb{P}_{k,2}$ and $\mathbb{S}_{k+1,2} = \boldsymbol{R}_{\frac{\pi}{2}}(\boldsymbol{x})\mathbb{P}^{\mathrm{H}}_{k,2}$.)

Exercise 15.5 (Hodge decomposition). Prove that for all $k \in \mathbb{N}$,

$$\mathbb{P}_{k+1,d} = \mathbb{N}_{k,d} \oplus \nabla\mathbb{P}^{\mathrm{H}}_{k+2,d}.$$

(*Hint*: compute $\mathbb{N}_{k,d} \cap \nabla\mathbb{P}^{\mathrm{H}}_{k+2,d}$, and use a dimension argument.)

Exercise 15.6 (Face element). We use the notation from the proof of Lemma 15.15. Let $F \in \mathcal{F}_K$. Let $\boldsymbol{T}_F : \widehat{S}^2 \to F$ be an affine bijective mapping. Let $\mathbb{J}_F$ be the Jacobian matrix of $\boldsymbol{T}_F$. Let $\boldsymbol{v} \in \mathbb{N}_{k,3}$ and let $\widehat{\boldsymbol{v}} := \mathbb{J}_F^{\mathsf{T}}(\mathbb{I}_3 - \boldsymbol{n}_F\otimes\boldsymbol{n}_F)(\boldsymbol{v}\circ\boldsymbol{T}_F)$. Show that $\widehat{\boldsymbol{v}} \in \mathbb{N}_{k,2}$. (*Hint*: compute $\widehat{\boldsymbol{y}}^{\mathsf{T}}\widehat{\boldsymbol{v}}(\widehat{\boldsymbol{y}})$ and apply the result from Exercise 14.4.)

Exercise 15.7 (Geometric mapping $\boldsymbol{T}_A$). Let A be an affine subspace of $\mathbb{R}^d$ of dimension $l \in \{1{:}d-1\}$, $d \geq 2$. Let $\boldsymbol{a} \in A$ and let $\boldsymbol{P}_A(\boldsymbol{x}) := \boldsymbol{a} + \Pi_A(\boldsymbol{x}-\boldsymbol{a})$ be the orthogonal projection onto A, where $\Pi_A \in \mathbb{R}^{d\times d}$. (i) Let $\boldsymbol{n} \in \mathbb{R}^d$ be such that $\boldsymbol{n}\cdot(\boldsymbol{x}-\boldsymbol{y}) = 0$ for all $\boldsymbol{x},\boldsymbol{y} \in A$ (we say that $\boldsymbol{n}$ is normal to A). Show that $\Pi_A\boldsymbol{n} = 0$. Let $\boldsymbol{t} \in \mathbb{R}^d$ be such that $\boldsymbol{a}+\boldsymbol{t} \in A$ (we say that $\boldsymbol{t}$ is tangent to A). Show that $\Pi_A(\boldsymbol{t}) = \boldsymbol{t}$. (ii) Let $q \in \mathbb{P}_{k,l}$ and let $\tilde{q}(\boldsymbol{x}) := q(\boldsymbol{T}_A^{-1}\circ\boldsymbol{P}_A(\boldsymbol{x}))$. Compute $\nabla\tilde{q}$. (iii) Show that there are $\boldsymbol{t}_1,\dots,\boldsymbol{t}_l$ tangent vectors and $q_1,\dots,q_l$ polynomials in $\mathbb{P}_{k,l}$ such that $\nabla\tilde{q}(\boldsymbol{x}) = \sum_{s\in\{1:l\}} q_s(\boldsymbol{T}_A^{-1}(\boldsymbol{x}))\boldsymbol{t}_s$ for all $\boldsymbol{x} \in A$. (iv) Let $\boldsymbol{t}$ be a tangent vector. Show that there is $\mu \in \mathbb{P}_{k,l}$ such that $\boldsymbol{t}\cdot\nabla\tilde{q}(\boldsymbol{x}) = \mu(\boldsymbol{T}_A^{-1}(\boldsymbol{x}))$.

Exercise 15.8 (Cartesian Nédélec element). (i) Propose a basis for $\mathbb{N}^{\square}_{0,3}$. (ii) Prove Proposition 15.23. (*Hint*: accept as a fact that any field $\boldsymbol{v} \in \mathbb{N}^{\square}_{k,3}$ annihiliating all the edge and faces dofs defined in (15.17) satisfies $\boldsymbol{v}_{|F}\times\boldsymbol{n}_F = \boldsymbol{0}$ for all $F \in \mathcal{F}_K$; then adapt the proof of Lemma 15.16 by using the $\mathbb{RT}^{\square}_{k,3}$ finite element defined in §14.5.2.)

Solution to exercises

Exercise 15.1 ($\mathbb{S}_{1,d}$). (i) Let $\boldsymbol{q} \in \mathbb{S}_{1,d}$. Since $\boldsymbol{q}$ is homogeneous of degree 1, there is a unique $d\times d$ matrix $\mathbb{Q}$ such $\boldsymbol{q}(\boldsymbol{x}) = \mathbb{Q}\boldsymbol{x}$. Then, $\boldsymbol{q} \in \mathbb{S}_{1,d}$ if and only if $\boldsymbol{x}^\mathsf{T}\mathbb{Q}\boldsymbol{x} = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$, which means that the quadratic form $\sum_{i\in\{1:d\}}\mathbb{Q}_{ii}x_i^2 + \sum_{i\neq j\in\{1:d\}}(\mathbb{Q}_{ij}+\mathbb{Q}_{ji})x_ix_j$ vanishes for all $\boldsymbol{x} \in \mathbb{R}^d$. Hence, $\mathbb{Q}$ is skew-symmetric. We have established that there is a one-to-one correspondence between the members of $\mathbb{S}_{1,d}$ and the $d\times d$ skew-symmetric matrices.
(ii) Consider the $\frac{d(d-1)}{2}$ skew-symmetric matrices $\mathbb{Q}^{ij}$, for all $i,j \in \{1{:}d\}$ with $i \neq j$, defined by $\mathbb{Q}^{ij}_{kl} := \delta_{ki}\delta_{lj} - \delta_{kj}\delta_{li}$ (the only nonzero entries of $\mathbb{Q}^{ij}$ are $\mathbb{Q}^{ij}_{ij} = 1$ and $\mathbb{Q}^{ij}_{ji} = -1$). Then, setting $\boldsymbol{q}^{ij}(\boldsymbol{x}) := \mathbb{Q}^{ij}\boldsymbol{x}$, we have shown that $\{\boldsymbol{q}^{ij}\}_{i,j\in\{1:d\},i\neq j}$ is a basis of $\mathbb{S}_{1,d}$.
(ii) Let us now focus on the case $d = 3$. The above definitions show that $\boldsymbol{q}^{12}(\boldsymbol{x}) = -\boldsymbol{e}_3\times\boldsymbol{x}$, $\boldsymbol{q}^{23}(\boldsymbol{x}) = -\boldsymbol{e}_1\times\boldsymbol{x}$ and $\boldsymbol{q}^{31}(\boldsymbol{x}) = -\boldsymbol{e}_2\times\boldsymbol{x}$. Hence, for all $\boldsymbol{q}(\boldsymbol{x}) = \beta_3\boldsymbol{q}^{12}(\boldsymbol{x}) + \beta_1\boldsymbol{q}^{23}(\boldsymbol{x}) + \beta_2\boldsymbol{q}^{31}(\boldsymbol{x}) \in \mathbb{S}_{1,3}$, we have $\boldsymbol{q}(\boldsymbol{x}) = \boldsymbol{b}\times\boldsymbol{x}$, where $\boldsymbol{b} := -\beta_1\boldsymbol{e}_1 - \beta_2\boldsymbol{e}_2 - \beta_3\boldsymbol{e}_3$.

Exercise 15.2 (Cross product). (i) Using Exercise 9.5, we obtain

$$(\mathbb{A}\boldsymbol{b})\times(\mathbb{A}\boldsymbol{c}) = \det(\mathbb{A})^{-1}\mathbb{A}^{-\mathsf{T}}(\boldsymbol{b}\times\boldsymbol{c}) = \mathbb{A}(\boldsymbol{b}\times\boldsymbol{c}),$$

since $\det(\mathbb{A}) = 1$ and $\mathbb{A}^{-\mathsf{T}} = \mathbb{A}$.
(ii) We have (using summation for repeated indices)

$$\begin{aligned}((\boldsymbol{a}\times\boldsymbol{b})\times\boldsymbol{c})_k &= -c_i\varepsilon_{ikj}a_l\varepsilon_{ljm}b_m\\ &= -c_ia_lb_m(\delta_{im}\delta_{kl} - \delta_{il}\delta_{km})\\ &= (\boldsymbol{a}\cdot\boldsymbol{c})b_k - (\boldsymbol{b}\cdot\boldsymbol{c})a_k,\end{aligned}$$

since $\varepsilon_{ikj}\varepsilon_{ljm} = \delta_{im}\delta_{kl} - \delta_{il}\delta_{km}$.
(iii) We apply the formula derived in Step (ii) using $\boldsymbol{n}\cdot\boldsymbol{n} = 1$.

Exercise 15.3 ($\mathbb{N}_{0,3}$). (i) Since $\boldsymbol{t}_E = \boldsymbol{z}_q - \boldsymbol{z}_p$, $(\boldsymbol{z}_q-\boldsymbol{z}_p)\cdot\nabla\lambda_q = \lambda_q(\boldsymbol{z}_q) - \lambda_q(\boldsymbol{z}_p) = 1$, and similarly $(\boldsymbol{z}_q-\boldsymbol{z}_p)\cdot\nabla\lambda_p = -1$, we infer that $\int_E \boldsymbol{\theta}^{\mathrm{e}}_E\cdot\boldsymbol{t}_E\,\mathrm{d}l = \int_E(\lambda_p+\lambda_q)\,\mathrm{d}l = |E|$ showing that $\sigma^{\mathrm{e}}_E(\boldsymbol{\theta}^{\mathrm{e}}_E) = 1$. Consider now an edge E' with $E' \neq E$. Then, at least one vertex $\boldsymbol{z}_p$ or $\boldsymbol{z}_q$ is not in E', say $\boldsymbol{z}_p \notin E'$. This implies that $\lambda_p = 0$ in E' and that $\boldsymbol{t}_{E'}\cdot\nabla\lambda_p = 0$. Hence, $\sigma^{\mathrm{e}}_{E'}(\boldsymbol{\theta}^{\mathrm{e}}_E) = 0$.
(ii) Let $\boldsymbol{v} \in \mathbb{N}_{0,3}$. Then $\boldsymbol{v} = \boldsymbol{a} + \boldsymbol{b}\times(\boldsymbol{x}-\boldsymbol{c}_K)$ with $\boldsymbol{a},\boldsymbol{b} \in \mathbb{R}^3$, and since $\boldsymbol{c}_K$ is the barycenter of K, we infer that $\boldsymbol{a} = \langle\boldsymbol{v}\rangle_K$. Furthermore, using the hint yields that $\nabla\times\boldsymbol{v} = 2\boldsymbol{b}$. In conclusion,

$\boldsymbol{v} = \langle \boldsymbol{v} \rangle_K + \frac{1}{2}(\nabla \times \boldsymbol{v}) \times (\boldsymbol{x} - \boldsymbol{c}_K)$.
(iii) The assertion $(\boldsymbol{\theta}_E^{\mathrm{e}})_{|F} \times \boldsymbol{n}_{K|F} = \boldsymbol{0}$ if E is not an edge of F is a direct consequence of Lemma 15.15 since the three dofs of $\boldsymbol{\theta}_E^{\mathrm{e}}$ attached to F vanish. Assume now that E is an edge of F. Observing that

$$\boldsymbol{\theta}_E^{\mathrm{e}} \times \boldsymbol{n}_{K|F} = \iota_{E,F} \boldsymbol{R}_{\frac{\pi}{2}} (\boldsymbol{\theta}_E^{\mathrm{e}} - (\boldsymbol{\theta}_E^{\mathrm{e}} \cdot \boldsymbol{n}_{K|F}) \boldsymbol{n}_{K|F}),$$

where $\boldsymbol{R}_{\frac{\pi}{2}}$ is the rotation by $\frac{\pi}{2}$ in the hyperplane parallel to F, and recalling that $\boldsymbol{R}_{\frac{\pi}{2}}(\boldsymbol{\theta}_E^{\mathrm{e}} - (\boldsymbol{\theta}_E^{\mathrm{e}} \cdot \boldsymbol{n}_{K|F}) \boldsymbol{n}_{K|F}) \circ \boldsymbol{T}_F$ is in $\mathbb{RT}_{0,2}$, we can use Exercise 14.1(ii) to infer that $\int_F \boldsymbol{\theta}_E^{\mathrm{e}} \times \boldsymbol{n}_{K|F} \, \mathrm{d}s = \iota_{E,F}(\boldsymbol{c}_E - \boldsymbol{c}_F)$.
(iv) Using the hint, we obtain that

$$\boldsymbol{e} \cdot \int_K \boldsymbol{\theta}_E^{\mathrm{e}} \, \mathrm{d}x = \int_K (\nabla \times \boldsymbol{\psi}) \cdot \boldsymbol{\theta}_E^{\mathrm{e}} = \int_K \boldsymbol{\psi} \cdot (\nabla \times \boldsymbol{\theta}_E^{\mathrm{e}}) \, \mathrm{d}x - \sum_{F \in \mathcal{F}_K} \int_F (\boldsymbol{n}_{K|F} \times \boldsymbol{\theta}_E^{\mathrm{e}}) \cdot \boldsymbol{\psi} \, \mathrm{d}s.$$

The first term on the right-hand side vanishes since $\boldsymbol{\psi}$ has zero mean value on K and $\nabla \times \boldsymbol{\theta}_E^{\mathrm{e}}$ is constant on K. Since the summation in the second term reduces to $F \in \mathcal{F}_E$ owing to Step (iii), we infer that

$$\boldsymbol{e} \cdot \int_K \boldsymbol{\theta}_E^{\mathrm{e}} \, \mathrm{d}x = - \sum_{F \in \mathcal{F}_E} \int_F (\boldsymbol{n}_{K|F} \times \boldsymbol{\theta}_E^{\mathrm{e}}) \cdot \boldsymbol{\psi} \, \mathrm{d}s =: \mathfrak{T}_1 + \mathfrak{T}_2,$$

with

$$\mathfrak{T}_1 := - \sum_{F \in \mathcal{F}_E} \int_F (\boldsymbol{n}_{K|F} \times \boldsymbol{\theta}_E^{\mathrm{e}}) \cdot \boldsymbol{\psi}(\boldsymbol{c}_F) \, \mathrm{d}s,$$

$$\mathfrak{T}_2 := - \sum_{F \in \mathcal{F}_E} \int_F (\boldsymbol{n}_{K|F} \times \boldsymbol{\theta}_E^{\mathrm{e}}) \cdot (\boldsymbol{\psi} - \boldsymbol{\psi}(\boldsymbol{c}_F)) \, \mathrm{d}s.$$

Since $\boldsymbol{\psi}(\boldsymbol{c}_F)$ is constant, we can use Step (iii) to evaluate $\mathfrak{T}_1$, so that

$$\begin{aligned} \mathfrak{T}_1 &= \frac{1}{2} \sum_{F \in \mathcal{F}_E} \iota_{E,F} (\boldsymbol{c}_E - \boldsymbol{c}_F) \cdot (\boldsymbol{e} \times (\boldsymbol{c}_F - \boldsymbol{c}_K)) \\ &= \frac{1}{2} \sum_{F \in \mathcal{F}_E} \iota_{E,F} \boldsymbol{e} \cdot ((\boldsymbol{c}_F - \boldsymbol{c}_K) \times (\boldsymbol{c}_E - \boldsymbol{c}_F)). \end{aligned}$$

Let us finally prove that $\mathfrak{T}_2 = \boldsymbol{0}$. Since $\boldsymbol{\psi} - \boldsymbol{\psi}(\boldsymbol{c}_F)$ has zero mean value on F, we can write

$$\mathfrak{T}_2 = - \sum_{F \in \mathcal{F}_E} \int_F (\boldsymbol{n}_{K|F} \times (\boldsymbol{\theta}_E^{\mathrm{e}} - \boldsymbol{\theta}_E^{\mathrm{e}}(\boldsymbol{c}_F))) \cdot (\boldsymbol{\psi} - \boldsymbol{\psi}(\boldsymbol{c}_F)) \, \mathrm{d}s - \sum_{F \in \mathcal{F}_E} \int_F I_F \, \mathrm{d}s.$$

Since $\boldsymbol{\theta}_E^{\mathrm{e}}(\boldsymbol{x}) - \boldsymbol{\theta}_E^{\mathrm{e}}(\boldsymbol{c}_F) = \boldsymbol{b} \times (\boldsymbol{x} - \boldsymbol{c}_F)$ for some $\boldsymbol{b} \in \mathbb{R}^3$ and since $\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}(\boldsymbol{c}_F) = \frac{1}{2} \boldsymbol{e} \times (\boldsymbol{x} - \boldsymbol{c}_F)$, we obtain that

$$\begin{aligned} I_F &:= \frac{1}{2} (\boldsymbol{n}_{K|F} \times (\boldsymbol{b} \times (\boldsymbol{x} - \boldsymbol{c}_F))) \cdot (\boldsymbol{e} \times (\boldsymbol{x} - \boldsymbol{c}_F)) \\ &= \frac{1}{2} (\boldsymbol{n}_{K|F} \cdot \boldsymbol{b}) (\boldsymbol{x} - \boldsymbol{c}_F) \cdot (\boldsymbol{e} \times (\boldsymbol{x} - \boldsymbol{c}_F)) = \boldsymbol{0}, \end{aligned}$$

since $\boldsymbol{n}_{K|F} \cdot (\boldsymbol{x} - \boldsymbol{c}_F) = 0$.

Exercise 15.4 (Rotated $\mathbb{RT}_{k,2}$). We observe that $\boldsymbol{R}_{\frac{\pi}{2}}\mathbb{P}_{k,2} = \mathbb{P}_{k,2}$. Moreover, we have

$$0 = \boldsymbol{x}\cdot\boldsymbol{q} = \sum_{l\in\{0:k+1\}} q_{1,l}x_1^{l+1}x_2^{k+1-l} + \sum_{l\in\{0:k+1\}} q_{2,l}x_1^{l}x_2^{k+2-l}.$$

This implies that $q_{1,k+1} = 0$, $q_{2,0} = 0$, and $q_{1,l} = -q_{2,l+1}$ for all $l \in \{0:k\}$. Hence, $q_1 = x_2 r$ and $q_2 = -x_1 r$ with $r = \sum_{l\in\{0:k\}} q_{1,l}x_1^l x_2^{k-l} \in \mathbb{P}^{\mathrm{H}}_{k,2}$. This shows that $\mathbb{S}_{k+1,2} = (\boldsymbol{R}_{\frac{\pi}{2}}\boldsymbol{x})\mathbb{P}^{\mathrm{H}}_{k,2}$. We conclude that $\mathbb{N}_{k,2} = \boldsymbol{R}_{\frac{\pi}{2}}(\mathbb{RT}_{k,2})$.

Exercise 15.5 (Hodge decomposition). Let $\boldsymbol{v} \in \mathbb{N}_{k,d} \cap \nabla\mathbb{P}^{\mathrm{H}}_{k+2,d}$, so that $\boldsymbol{v} = \nabla p$ where $p \in \mathbb{P}^{\mathrm{H}}_{k+2,d}$. Observe that $\nabla p \in \mathbb{P}^{\mathrm{H}}_{k+1,d}$ and $\boldsymbol{x}\cdot\nabla p(\boldsymbol{x}) = (k+2)p(\boldsymbol{x})$. The assumption $\boldsymbol{v} = \nabla p \in \mathbb{N}_{k,d}$ and the property $\nabla p \in \mathbb{P}^{\mathrm{H}}_{k+1,d}$ imply that $\boldsymbol{x}\cdot\nabla p(\boldsymbol{x}) = 0$, which can be true only if $p = 0$. Hence, $\boldsymbol{v} = \boldsymbol{0}$. We conclude by using a dimension argument, since we have

$$\begin{aligned}\dim(\mathbb{N}_{k,d}) + \dim(\nabla\mathbb{P}^{\mathrm{H}}_{k+2,d}) &= \dim(\mathbb{N}_{k,d}) + \dim(\mathbb{P}^{\mathrm{H}}_{k+2,d})\\ &= \frac{(k+d+1)!}{k!(d-1)!(k+2)} + \frac{(k+d+1)!}{(k+2)!(d-1)!}\\ &= d\frac{(k+d+1)!}{(k+1)!d!} = \dim(\mathbb{P}_{k+1,d}).\end{aligned}$$

Exercise 15.6 (Face element). By definition, we have $\boldsymbol{v} = \boldsymbol{r} + \boldsymbol{q}$ where $\boldsymbol{r} \in \mathbb{P}_{k,3}$ and $\boldsymbol{q} \in \mathbb{P}^{\mathrm{H}}_{k+1,3}$ satisfies $\boldsymbol{y}^{\mathsf{T}}\boldsymbol{q}(\boldsymbol{y}) = 0$. Let $\Pi_F := \mathbb{I}_3 - \boldsymbol{n}_F\otimes\boldsymbol{n}_F$. Let $\widehat{\boldsymbol{y}} \in \mathbb{R}^2$. We have

$$\begin{aligned}\widehat{\boldsymbol{y}}^{\mathsf{T}}\widehat{\boldsymbol{v}}(\widehat{\boldsymbol{y}}) &= \widehat{\boldsymbol{y}}^{\mathsf{T}}\mathbb{J}_F^{\mathsf{T}}\Pi_F(\boldsymbol{v}\circ\boldsymbol{T}_F)(\widehat{\boldsymbol{y}})\\ &= \widehat{\boldsymbol{y}}^{\mathsf{T}}\mathbb{J}_F^{\mathsf{T}}\Pi_F\boldsymbol{r}(\boldsymbol{T}_F(\widehat{\boldsymbol{y}})) + (\mathbb{J}_F\widehat{\boldsymbol{y}})^{\mathsf{T}}\Pi_F\boldsymbol{q}(\boldsymbol{T}_F(\widehat{\boldsymbol{y}}))\\ &= \widehat{\boldsymbol{y}}^{\mathsf{T}}\mathbb{J}_F^{\mathsf{T}}\Pi_F\boldsymbol{r}(\boldsymbol{T}_F(\widehat{\boldsymbol{y}})) + (\mathbb{J}_F\widehat{\boldsymbol{y}})^{\mathsf{T}}\Pi_F\boldsymbol{q}(\boldsymbol{T}_F(\widehat{\boldsymbol{y}}) - \boldsymbol{T}_F(\boldsymbol{0}_{\mathbb{R}^2}) + \boldsymbol{T}_F(\boldsymbol{0}_{\mathbb{R}^2}))\\ &= \widehat{\boldsymbol{y}}^{\mathsf{T}}\mathbb{J}_F^{\mathsf{T}}\Pi_F\boldsymbol{r}(\boldsymbol{T}_F(\widehat{\boldsymbol{y}})) + (\mathbb{J}_F\widehat{\boldsymbol{y}})^{\mathsf{T}}\Pi_F\boldsymbol{q}(\mathbb{J}_F(\widehat{\boldsymbol{y}}) + \boldsymbol{T}_F(\boldsymbol{0}_{\mathbb{R}^2})).\end{aligned}$$

We now invoke the result from Exercise 14.4 componentwise: there is $\widehat{\boldsymbol{t}} \in \mathbb{P}_{k,2}$ such that $\boldsymbol{q}(\mathbb{J}_F(\widehat{\boldsymbol{y}}) + \boldsymbol{T}_F(\boldsymbol{0}_{\mathbb{R}^2})) = \boldsymbol{q}(\mathbb{J}_F(\widehat{\boldsymbol{y}})) + \widehat{\boldsymbol{t}}(\widehat{\boldsymbol{y}})$. Setting $\widehat{s}(\widehat{\boldsymbol{y}}) := \widehat{\boldsymbol{y}}^{\mathsf{T}}\mathbb{J}_F^{\mathsf{T}}\Pi_F(\boldsymbol{r}(\boldsymbol{T}_F(\widehat{\boldsymbol{y}})) + \widehat{\boldsymbol{t}}(\widehat{\boldsymbol{y}}))$ where $\widehat{s} \in \mathbb{P}_{k+1,2}$, and observing that $(\mathbb{J}_F\widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{n}_F = 0$ for all $\widehat{\boldsymbol{y}}$, we obtain

$$\widehat{\boldsymbol{y}}^{\mathsf{T}}\widehat{\boldsymbol{v}}(\widehat{\boldsymbol{y}}) = \widehat{s}(\boldsymbol{y}) + (\mathbb{J}_F\widehat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{q}(\mathbb{J}_F(\widehat{\boldsymbol{y}})) = \widehat{s}(\boldsymbol{y}).$$

Since $\widehat{\boldsymbol{v}} \in \mathbb{P}_{k+1,2}$, we have the decomposition $\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{r}} + \widehat{\boldsymbol{q}}$ where $\widehat{\boldsymbol{r}} \in \mathbb{P}_{k,2}$ and $\widehat{\boldsymbol{q}} \in \mathbb{P}^{\mathrm{H}}_{k+1,2}$. We have

$$\widehat{\boldsymbol{y}}^{\mathsf{T}}\widehat{\boldsymbol{v}}(\widehat{\boldsymbol{y}}) = \widehat{\boldsymbol{y}}^{\mathsf{T}}\widehat{\boldsymbol{r}}(\widehat{\boldsymbol{y}}) + \widehat{\boldsymbol{y}}^{\mathsf{T}}\widehat{\boldsymbol{q}}(\widehat{\boldsymbol{y}}) = \widehat{s}(\boldsymbol{y}) \in \mathbb{P}_{k+1,2},$$

but $\widehat{\boldsymbol{y}}^{\mathsf{T}}\widehat{\boldsymbol{r}}(\widehat{\boldsymbol{y}}) \in \mathbb{P}_{k+1,2}$ and $\widehat{\boldsymbol{y}}^{\mathsf{T}}\widehat{\boldsymbol{q}}(\widehat{\boldsymbol{y}}) \in \mathbb{P}^{\mathrm{H}}_{k+2,2}$. Hence, $\widehat{\boldsymbol{y}}^{\mathsf{T}}\widehat{\boldsymbol{q}}(\widehat{\boldsymbol{y}}) = 0$ for all $\widehat{\boldsymbol{y}}$, which proves that $\widehat{\boldsymbol{v}} \in \mathbb{N}_{k,2}$.

Exercise 15.7 (Geometric mapping $\boldsymbol{T}_A$). (i) These are elementary results in linear algebra. Let $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_l$ be a basis of $A - \boldsymbol{a}$. Let $\boldsymbol{n}_{l+1}, \ldots, \boldsymbol{n}_d$ be a basis of $\mathrm{span}\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_l\}^{\perp}$. Let $\boldsymbol{n}$ be a normal vector. Let $\boldsymbol{x} := \boldsymbol{a}+\boldsymbol{n}$. Then $\boldsymbol{P}_A(\boldsymbol{x})-\boldsymbol{a} = \Pi_A(\boldsymbol{x}-\boldsymbol{a}) = \Pi_A(\boldsymbol{n})$. Observe that $0 = \boldsymbol{n}_s\cdot(\boldsymbol{P}_A(\boldsymbol{x})-\boldsymbol{a}) = \boldsymbol{n}_s\cdot\Pi_A(\boldsymbol{n})$, for every normal vector $\boldsymbol{n}_s$ for all $s \in \{l+1:d\}$. Note also that $\boldsymbol{t}_s\cdot(\boldsymbol{P}_A(\boldsymbol{x}) - \boldsymbol{x}) = 0$ for every tangent vector $\boldsymbol{t}_s$ for all $s \in \{1:d\}$, i.e., $0 = \boldsymbol{t}_s\cdot(\boldsymbol{a} + \Pi_A(\boldsymbol{n}) - \boldsymbol{x}) = \boldsymbol{t}_s\cdot(-\boldsymbol{n} + \Pi_A(\boldsymbol{n})) = \boldsymbol{t}_s\cdot\Pi_A(\boldsymbol{n})$. Hence, $\Pi_A(\boldsymbol{n})$ is orthogonal to $\mathrm{span}\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_l\} \oplus \mathrm{span}\{\boldsymbol{n}_{l+1}, \ldots, \boldsymbol{n}_d\} = \mathbb{R}^d$, meaning that $\Pi_A(\boldsymbol{n}) = \boldsymbol{0}$. Let $\boldsymbol{t}$ be a tangent vector and let $\boldsymbol{x} = \boldsymbol{a} + \boldsymbol{t}$, so that $\boldsymbol{x} \in A$ by definition. Hence,

$\boldsymbol{t} = \boldsymbol{x} - \boldsymbol{a} = \boldsymbol{P}_A(\boldsymbol{x}) - \boldsymbol{a} = \Pi_A(\boldsymbol{x} - \boldsymbol{a}) = \Pi_A(\boldsymbol{t})$.
(ii) Let $\boldsymbol{h} \in \mathbb{R}^d$. We use the Fréchet derivative notation and apply the chain rule. This gives

$$\begin{aligned} D\tilde{q}(\boldsymbol{x})(\boldsymbol{h}) &= Dq(\boldsymbol{T}_A^{-1} \circ \boldsymbol{P}_A(\boldsymbol{x}))(D(\boldsymbol{T}_A^{-1} \circ \boldsymbol{P}_A(\boldsymbol{x}))(\boldsymbol{h})) \\ &= Dq(\boldsymbol{T}_A^{-1} \circ \boldsymbol{P}_A(\boldsymbol{x}))(D\boldsymbol{T}_A^{-1}(\boldsymbol{P}_A(\boldsymbol{x}))(D\boldsymbol{P}_A(\boldsymbol{x})(\boldsymbol{h}))). \end{aligned}$$

Note that $D\boldsymbol{P}_A(\boldsymbol{x})(\boldsymbol{h}) = \Pi_A(\boldsymbol{h})$ and $D\boldsymbol{T}_A^{-1}(\boldsymbol{x}')(\boldsymbol{h}') = \mathbb{J}_A^{-1}\boldsymbol{h}'$ for all $\boldsymbol{x}' \in A$ and all $\boldsymbol{h}' \in \mathbb{R}^d$. We identify the Fréchet derivatives of $\tilde{q}$ and q with the gradients, so that

$$\begin{aligned} \nabla\tilde{q}(\boldsymbol{x})\cdot\boldsymbol{h} := D\tilde{q}(\boldsymbol{x})(\boldsymbol{h}) &= \nabla q(\boldsymbol{T}_A^{-1} \circ \boldsymbol{P}_A(\boldsymbol{x}))\cdot(\mathbb{J}_A^{-1}\Pi_A(\boldsymbol{h})) \\ &= \Pi_A(\mathbb{J}_A^{-\mathsf{T}}\nabla q(\boldsymbol{T}_A^{-1} \circ \boldsymbol{P}_A(\boldsymbol{x})))\cdot\boldsymbol{h}, \qquad \forall \boldsymbol{h} \in \mathbb{R}^d. \end{aligned}$$

Hence, we have

$$\nabla\tilde{q}(\boldsymbol{x}) = \Pi_A^{\mathsf{T}}(\mathbb{J}_A^{-\mathsf{T}}\nabla q(\boldsymbol{T}_A^{-1} \circ \boldsymbol{P}_A(\boldsymbol{x}))), \qquad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

(iii) Let $\boldsymbol{n}$ any normal vector. We have

$$\boldsymbol{n}\cdot\nabla\tilde{q}(\boldsymbol{x}) = \boldsymbol{n}\cdot\Pi_A^{\mathsf{T}}(\mathbb{J}_A^{-\mathsf{T}}\nabla q(\boldsymbol{T}_A^{-1} \circ \boldsymbol{P}_A(\boldsymbol{x}))) = \Pi_A(\boldsymbol{n})\cdot\mathbb{J}_A^{-\mathsf{T}}\nabla q(\boldsymbol{T}_A^{-1} \circ \boldsymbol{P}_A(\boldsymbol{x})) = 0,$$

since we have already proved that $\Pi_A(\boldsymbol{n}) = \boldsymbol{0}$. Hence, $\nabla\tilde{q}(\boldsymbol{x}) \in \operatorname{span}\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_l\}$. Moreover, since $P_A(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in A$, we have

$$\nabla\tilde{q}(\boldsymbol{x}) = \Pi_A(\mathbb{J}_A^{-\mathsf{T}}\nabla q(\boldsymbol{T}_A^{-1}(\boldsymbol{x}))), \qquad \forall \boldsymbol{x} \in A.$$

Hence, $\nabla\tilde{q}(\boldsymbol{x})$ is an l-variate $\mathbb{R}^d$-valued polynomial of degree at most k. The above two arguments show that there exist $q_1, \ldots, q_l \in \mathbb{P}_{l,d}$ such that

$$\nabla\tilde{q}(\boldsymbol{x}) = \sum_{s\in\{1:l\}} q_s(\boldsymbol{T}_A^{-1}(\boldsymbol{x}))\boldsymbol{t}_s, \qquad \forall \boldsymbol{x} \in A.$$

(iv) Let $\boldsymbol{t}$ be a tangent vector. The above arguments show that there is $\mu \in \mathbb{P}_{k,l}$ such that

$$\boldsymbol{t}\cdot\nabla\tilde{q}(\boldsymbol{x}) = \mu(\boldsymbol{T}_A^{-1}(\boldsymbol{x})), \qquad \forall \boldsymbol{x} \in A.$$

Exercise 15.8 (Cartesian Nédélec element). (i) A basis for $\mathbb{N}_{0,3}^{\square}$ is

$$\left\{ \begin{pmatrix} x_2x_3 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} x_2(1-x_3) \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (1-x_2)x_3 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (1-x_2)(1-x_3) \\ 0 \\ 0 \end{pmatrix}, \right.$$
$$\begin{pmatrix} 0 \\ x_3x_1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ x_3(1-x_1) \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ (1-x_3)x_1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ (1-x_3)(1-x_1) \\ 0 \end{pmatrix},$$
$$\left. \begin{pmatrix} 0 \\ 0 \\ x_1x_2 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ x_1(1-x_2) \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ (1-x_1)x_2 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ (1-x_1)(1-x_2) \end{pmatrix} \right\}.$$

(ii) Observe first that $\operatorname{card}(\Sigma) = 3k^2(k+1) + 12k(k+1) + 12(k+1) = 3(k+1)(k+2)^2 = \dim\mathbb{N}_{k,3}^{\square}$. It remains to show that any field $\boldsymbol{v} \in \mathbb{N}_{k,3}^{\square}$ that annihilates all the dofs defined in (15.17) vanishes identically. Owing to the hint, we already know that $\boldsymbol{v}_{|\partial K}\times\boldsymbol{n}_K = \boldsymbol{0}$.
(ii.a) Let us first prove that $\boldsymbol{w} := \nabla\times\boldsymbol{v} = \boldsymbol{0}$. The definition of the polynomial spaces $\mathbb{N}_{k,3}^{\square}$ and $\mathbb{RT}_{k,3}^{\square}$ implies that $\boldsymbol{w} \in \mathbb{Q}_{k+1,k,k}\times\mathbb{Q}_{k,k+1,k}\times\mathbb{Q}_{k,k,k+1} = \mathbb{RT}_{k,3}^{\square}$. Thus, we are going to show that $\boldsymbol{w} = \boldsymbol{0}$ by invoking Proposition 14.24, i.e., by showing that $\boldsymbol{w}$ annihilates all the dofs defined in (14.16). First, we observe that the normal component of $\boldsymbol{w}$ vanishes on ∂K since $\boldsymbol{v}_{|\partial K}\times\boldsymbol{n}_K = \boldsymbol{0}$. Therefore, $\boldsymbol{w}$ annihilates all the face dofs in (14.16). Let $j \in \{1{:}d\}$ and let $\{\psi_{j,m}\}_{m\in\{1:n_{\mathrm{sh}}^{\mathrm{c}}\}}$

be a basis of $\mathbb{Q}_{\alpha_1,\alpha_2,\alpha_3}$ with $\alpha_j = k-1$ and $\alpha_{j'} = k$ if $j' \neq j$. Let $\{\boldsymbol{\nu}_{K,j} := |F_j|\boldsymbol{e}_j\}_{j\in\{1:d\}}$ be the vectors orienting K, where $\{\boldsymbol{e}_j\}_{j\in\{1:d\}}$ is the canonical Cartesian basis of $\mathbb{R}^d$. Setting $\boldsymbol{\phi}_{j,m} := \boldsymbol{\nu}_{K,j}\psi_{m,j}$, we have $\int_K \boldsymbol{w}\cdot\boldsymbol{\phi}_{j,m}\,\mathrm{d}x = \int_K \boldsymbol{v}\cdot\nabla\times\boldsymbol{\phi}_{j,m}\,\mathrm{d}x$ since $\boldsymbol{v}_{|\partial K}\times\boldsymbol{n}_K = \boldsymbol{0}$. Since we have $\nabla\times\boldsymbol{\phi}_{j,m} = \sum_{j\in\{1:d\}}\boldsymbol{e}_j r_{j,m}$ with $r_{j,m}\in\mathbb{Q}_{\beta_1,\beta_2,\beta_3}$ with $\beta_j = k$ and $\beta_{j'} = k-1$ if $j'\neq j$, we infer from (15.17c) that $\int_K \boldsymbol{v}\cdot\nabla\times\boldsymbol{\phi}_{j,m}\,\mathrm{d}x = 0$. In conclusion, $\boldsymbol{w}$ annihilates all the cell dofs of the $\mathbb{RT}^{\square}_{k,3}$ finite element as well. Hence, $\boldsymbol{w} = \boldsymbol{0}$.

(ii.b) Since the field $\boldsymbol{v}\in\mathbb{N}^{\square}_{k,3}$ is curl-free, there is $q\in\mathbb{Q}_{k+1,3}$ such that $\boldsymbol{v} = \nabla q$. The property $\boldsymbol{v}_{|\partial K}\times\boldsymbol{n}_K = \boldsymbol{0}$ implies that q is constant on ∂K, and without loss of generality, we assume that $q_{|\partial K} = 0$. If $k = 0$, this implies that $q = 0$, so that it remains to consider the case $k\geq 1$. In this situation, there is $\tilde{q}\in\mathbb{Q}_{k-1,3}$ such that $q = b\tilde{q}$ with $b := \prod_{i\in\{1:d\}} x_i(1-x_i)$. Let us write $\tilde{q} = \sum_{\alpha\in\mathcal{B}_{k-1,3}} a_\alpha \boldsymbol{x}^\alpha$ with $\mathcal{B}_{k-1,3} := \{\alpha\in\mathbb{N}^3 \mid \alpha_i\in\{0{:}k{-}1\},\ \forall i\in\{1{:}d\}\}$. We consider the polynomial $r(\boldsymbol{x}) := \sum_{\alpha\in\mathcal{B}_{k-1,3}} \frac{1}{\alpha_1+1} a_\alpha x_1 \boldsymbol{x}^\alpha$. We have $r\in\mathbb{Q}_{k,k-1,k-1}$ so that (15.17c) implies that $\int_K \boldsymbol{v}\cdot(\boldsymbol{e}_1 r)\,\mathrm{d}x = 0$. Since $\boldsymbol{v} = \nabla q$, $\nabla\cdot(\boldsymbol{e}_1 r) = \tilde{q}$, and $q_{|\partial K} = 0$, we infer that

$$0 = \int_K \boldsymbol{v}\cdot(\boldsymbol{e}_1 r)\,\mathrm{d}x = -\int_K b\tilde{q}^2\,\mathrm{d}x,$$

which proves that $\tilde{q}$ vanishes identically. In conclusion, we have shown that $\boldsymbol{v} = \boldsymbol{0}$. This completes the proof.

Chapter 16

Local interpolation in $\boldsymbol{H}(\mathrm{div})$ and $\boldsymbol{H}(\mathrm{curl})$ (I)

Exercises

Exercise 16.1 ($\check{\boldsymbol{V}}^{\mathrm{d}}(K)$). Show that $\boldsymbol{V}^{\mathrm{d}}(K)$ defined in (16.2) can be used in the commuting diagram of Lemma 16.2 after replacing $L^1(K)$ by $W^{s-1,p}(K)$. (*Hint*: use Theorem 3.19.)

Exercise 16.2 ($\mathcal{I}_K^{\mathrm{d}}$). Prove that the estimate (16.6) holds true for all $r \in [1, k+1]$, $r \notin \mathbb{N}$, every integer $m \in \{0{:}\lfloor r \rfloor\}$, and all $p \in [1, \infty)$. Prove that (16.7) holds true for all $r \in [0, k+1]$, $r \notin \mathbb{N}$, every integer $m \in \{0{:}\lfloor r \rfloor\}$, and all $p \in [1, \infty)$. (*Hint*: combine $W^{m,p}$-stability with Corollary 12.13.)

Exercise 16.3 (de Rham). Prove that the leftmost diagram in Lemma 16.16 commutes. (*Hint*: verify that $\nabla \mathcal{I}_K^{\mathrm{g}}(v) - \mathcal{I}_K^{\mathrm{c}}(\nabla v)$ annihilates all dofs in $\mathbb{N}_{k,d}$.)

Exercise 16.4 (Poincaré operators). Assume that K is star-shaped with respect to a point $\boldsymbol{a} \in K$. Let f and $\boldsymbol{g}$ be smooth functions on K. Define $P^{\mathrm{g}}(\boldsymbol{g})(\boldsymbol{x}) := (\boldsymbol{x}-\boldsymbol{a}) \cdot \int_0^1 \boldsymbol{g}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))\,\mathrm{d}t$, $\boldsymbol{P}^{\mathrm{c}}(\boldsymbol{g})(\boldsymbol{x}) := -(\boldsymbol{x}-\boldsymbol{a}) \times \int_0^1 \boldsymbol{g}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))\,\mathrm{d}t$ (if $d=3$), and $\boldsymbol{P}^{\mathrm{d}}(f)(\boldsymbol{x}) := (\boldsymbol{x}-\boldsymbol{a}) \int_0^1 f(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))t^{d-1}\,\mathrm{d}t$. Verify that (i) $\nabla P^{\mathrm{g}}(\boldsymbol{g}) = \boldsymbol{g}$ if $\partial_i g_j = \partial_j g_i$ for all $i,j \in \{1{:}d\}$; (ii) $\nabla \times \boldsymbol{P}^{\mathrm{c}}(\boldsymbol{g}) = \boldsymbol{g}$ if $\nabla \cdot \boldsymbol{g} = 0$; (iii) $\nabla \cdot \boldsymbol{P}^{\mathrm{d}}(f) = f$.

Exercise 16.5 (Koszul operator). (i) Let $\boldsymbol{v} \in \mathbb{P}_{k,d}^{\mathrm{H}}$ with $d=3$. Prove that $\nabla(\boldsymbol{x}\cdot\boldsymbol{v}) - \boldsymbol{x}\times(\nabla\times\boldsymbol{v}) = (k+1)\boldsymbol{v}$ and $-\nabla\times(\boldsymbol{x}\times\boldsymbol{v}) + \boldsymbol{x}(\nabla\cdot\boldsymbol{v}) = (k+2)\boldsymbol{v}$. (*Hint*: use Euler's identity from Lemma 14.3.) (ii) Prove that $\mathbb{P}_{k,d} = \nabla\mathbb{P}_{k+1,d} \oplus (\boldsymbol{x}\times\mathbb{P}_{k-1,d}) = \nabla\times\mathbb{P}_{k+1,d} \oplus (\boldsymbol{x}\mathbb{P}_{k-1,d})$. (*Hint*: establish first these identities for homogeneous polynomials.) *Note*: defining the Koszul operators $\kappa^{\mathrm{g}}(\boldsymbol{v}) := \boldsymbol{x}\cdot\boldsymbol{v}$ and $\boldsymbol{\kappa}^{\mathrm{c}}(\boldsymbol{v}) := -\boldsymbol{x}\times\boldsymbol{v}$ for vector fields and $\boldsymbol{\kappa}^{\mathrm{d}}(v) := \boldsymbol{x}v$ for scalar fields, one has $\kappa^{\mathrm{g}}(\nabla q) = kq$ (Euler's identity) and $\nabla\cdot(\boldsymbol{\kappa}^{\mathrm{d}}(q)) = (k+d)q$ for all $q \in \mathbb{P}_{k,d}^{\mathrm{H}}$, and $\nabla(\kappa^{\mathrm{g}}(\boldsymbol{q})) + \boldsymbol{\kappa}^{\mathrm{c}}(\nabla\times\boldsymbol{q}) = (k+1)\boldsymbol{q}$ and $\nabla\times(\boldsymbol{\kappa}^{\mathrm{c}}(\boldsymbol{q})) + \boldsymbol{\kappa}^{\mathrm{d}}(\nabla\cdot\boldsymbol{q}) = (k+2)\boldsymbol{q}$ for all $\boldsymbol{q} \in \mathbb{P}_{k,d}^{\mathrm{H}}$; see [1, Sec. 3.2].

Exercise 16.6 ($\nabla\cdot\mathbb{RT}_{k,d}$ and $\nabla\times\mathbb{N}_{k,3}$). (i) Prove that $\nabla\cdot\mathbb{RT}_{k,d} = \mathbb{P}_{k,d}$. (*Hint*: prove that $\nabla\cdot : \boldsymbol{x}\mathbb{P}_{k,d} \to \mathbb{P}_{k,d}$ is injective using Lemma 14.3.) (ii) Let us set $\mathbb{RT}_{k,d}^{\mathrm{div}=0} := \{\boldsymbol{v} \in \mathbb{RT}_{k,d} \mid \nabla\cdot\boldsymbol{v} = 0\}$. Determine $\dim(\mathbb{RT}_{k,d}^{\mathrm{div}=0})$ for $d \in \{2,3\}$. (iii) Show that $\mathbb{RT}_{k,3}^{\mathrm{div}=0} = \nabla\times\mathbb{P}_{k+1,3}$. (*Hint*: use Lemma 14.9.) (iv) Prove that $\mathbb{RT}_{k,3}^{\mathrm{div}=0} = \nabla\times\mathbb{N}_{k,3}$. (*Hint*: use the rank nullity theorem.)

Exercise 16.7 ($\nabla\mathbb{P}_{k+1,d}$ and $\nabla\times\boldsymbol{\mathbb{P}}_{k+1,3}$). Let $k\in\mathbb{N}$. (i) Set $\mathbb{P}^{\mathrm{c}}_{k,d}:=\nabla\mathbb{P}_{k+1,d}$. Show that $\dim(\mathbb{P}^{\mathrm{c}}_{k,d})=\binom{k+d+1}{d}-1$. (ii) Assume $d=3$. Set $\mathbb{P}^{\mathrm{d}}_{k,3}:=\nabla\times\boldsymbol{\mathbb{P}}_{k+1,3}$. Show that $\dim(\mathbb{P}^{\mathrm{d}}_{k,3})=3\binom{k+4}{3}-\binom{k+5}{3}+1=3\binom{k+3}{3}-\binom{k+2}{3}$ (with the convention that $\binom{2}{3}=0$). (*Hint*: use the exact cochain complex $\mathbb{P}_{0,d}\xrightarrow{i}\mathbb{P}_{k+2,d}\xrightarrow{\nabla}\boldsymbol{\mathbb{P}}_{k+1,d}\xrightarrow{\nabla\times}\boldsymbol{\mathbb{P}}_{k,d}\xrightarrow{\nabla\cdot}\mathbb{P}_{k-1,d}\xrightarrow{o}\{0\}$.)

Solution to exercises

Exercise 16.1 ($\check{\boldsymbol{V}}^{\mathrm{d}}(K)$). Let $s<1$ (the case $s\geq 1$ is trivial). The proof of Proposition 16.1 shows that after integration by parts, the term $\int_K(\nabla\cdot\boldsymbol{v})q\,\mathrm{d}x$ can be given a weak meaning for $\boldsymbol{v}\in\boldsymbol{V}^{\mathrm{d}}(K)$ and $q\in\mathbb{P}_{k,d}$. One replaces $L^1(K)$ by $W^{s-1,p}(K):=(W_0^{1-s,p'}(K))'$ and extends the domain of $\mathcal{I}_K^{\mathrm{b}}$ to $W^{s-1,p}(K)$, which is legitimate since $W_0^{1-s,p'}(K)=W^{1-s,p'}(K)$ owing to (3.5a) and $1-s<1-\frac{1}{p}=\frac{1}{p'}$.

Exercise 16.2 ($\mathcal{I}_K^{\mathrm{d}}$). Let us consider the estimate (16.6). Let $r\in[1,k+1]$, $r\notin\mathbb{N}$, $m\in\{0\!:\lfloor r\rfloor\}$, and $p\in[1,\infty)$. Notice that $\lfloor r\rfloor\geq 1$. For all $m\in\{1\!:\lfloor r\rfloor\}$, we have $W^{m,p}(K)\hookrightarrow\boldsymbol{V}^{\mathrm{d}}(K)$ (see (16.2)). Hence, $\mathcal{I}_K^{\mathrm{d}}$ is $W^{m,p}$-stable. Since $m\leq\lfloor r\rfloor\leq k$ and since $\mathbb{P}_{k,d}\subset\mathbb{RT}_{k,d}$ is pointwise invariant under $\mathcal{I}_K^{\mathrm{d}}$, we infer that

$$|\boldsymbol{v}-\mathcal{I}_K^{\mathrm{d}}(\boldsymbol{v})|_{\boldsymbol{W}^{m,p}(K)}\leq c\inf_{\boldsymbol{q}\in\boldsymbol{\mathbb{P}}_{k,d}}|\boldsymbol{v}-\boldsymbol{q}|_{\boldsymbol{W}^{m,p}(K)},$$

and we conclude by invoking Corollary 12.13. If $m=0$, we reason similarly by using the fact that the stability property (16.8) also holds true for $r\geq 1$ (because $W^{1,p}(K)\hookrightarrow\boldsymbol{V}^{\mathrm{d}}(K)$), and we conclude as above. Finally, the reasoning for the estimate on the divergence is similar since Lemma 11.18 implies that $\mathcal{I}_K^{\mathrm{b}}$ is $W^{m,p}$-stable for all $m\in\{0\!:\lfloor r\rfloor\}$.

Exercise 16.3 (de Rham). We first notice that $\nabla\mathbb{P}_{k+1,d}\subset\mathbb{P}_{k,d}\subset\mathbb{N}_{k,d}$. Let us prove that $\sigma_i(\boldsymbol{\delta})=0$ with $\boldsymbol{\delta}=\nabla\mathcal{I}_K^{\mathrm{g}}(v)-\mathcal{I}_K^{\mathrm{c}}(\nabla v)$ for all $v\in\check{V}^{\mathrm{g}}(K)$ and all the dofs $\{\sigma_i\}_{i\in\mathcal{N}}$ of the $\mathbb{N}_{k,d}$ element. Let $E\in\mathcal{E}_K$ be an edge of K with geometric mapping $\boldsymbol{T}_E$ and let $\boldsymbol{z}_p$, $\boldsymbol{z}_q$ be the end vertices of E such that $\boldsymbol{t}_E=\boldsymbol{z}_q-\boldsymbol{z}_p$. Set $\boldsymbol{\tau}_E:=|E|^{-1}\boldsymbol{t}_E$. Let $\mu_m\in\mathbb{P}_{k,1}$. Using $\boldsymbol{\tau}_E\cdot\nabla(\mu_m\circ\boldsymbol{T}_E^{-1})=\mu_m'\circ\boldsymbol{T}_E^{-1}$ and $\mu_m'\in\mathbb{P}_{k-1,1}$ (if $k\geq 1$, or $\mu_m'=0$ otherwise), together with the definitions of the dofs (7.11a), (7.11b), and (15.8a), we infer that (denoting $\mu_{E,m}:=\mu_m\circ\boldsymbol{T}_E^{-1}$)

$$\begin{aligned}\int_E\nabla\mathcal{I}_K^{\mathrm{g}}(v)\cdot\boldsymbol{\tau}_E\mu_{E,m}\,\mathrm{d}l&=\Big[\mathcal{I}_K^{\mathrm{g}}(v)\mu_{E,m}\Big]_{\boldsymbol{z}_p}^{\boldsymbol{z}_q}-\int_E\boldsymbol{\tau}_E\cdot\nabla(\mu_m\circ\boldsymbol{T}_E^{-1})\mathcal{I}_K^{\mathrm{g}}(v)\,\mathrm{d}l\\&=\Big[v\mu_{E,m}\Big]_{\boldsymbol{z}_p}^{\boldsymbol{z}_q}-\int_E(\mu_m'\circ\boldsymbol{T}_E^{-1})v\,\mathrm{d}l\\&=\int_E\nabla v\cdot\boldsymbol{\tau}_E\mu_{E,m}\,\mathrm{d}l=\int_E\mathcal{I}_K^{\mathrm{c}}(\nabla v)\cdot\boldsymbol{\tau}_E\mu_{E,m}\,\mathrm{d}l.\end{aligned}$$

Hence, $\boldsymbol{\delta}$ annihilates all the edge dofs of the $\mathbb{N}_{k,d}$ element. The proof is similar for the surface and volume dofs.

Exercise 16.4 (Poincaré operators). (i) We have

$$\begin{aligned}\nabla P^{\mathrm{g}}(\boldsymbol{g})(\boldsymbol{x}) &= \int_0^1 \boldsymbol{g}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))\,\mathrm{d}t + \int_0^1 \nabla\boldsymbol{g}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))\cdot(\boldsymbol{x}-\boldsymbol{a})t\,\mathrm{d}t\\ &= \boldsymbol{g}(\boldsymbol{x}) - \int_0^1 \tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))t\,\mathrm{d}t + \int_0^1 \nabla\boldsymbol{g}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))\cdot(\boldsymbol{x}-\boldsymbol{a})t\,\mathrm{d}t\\ &= \boldsymbol{g}(\boldsymbol{x}) + \int_0^1 (\nabla\boldsymbol{g}-\nabla\boldsymbol{g}^{\mathsf{T}})(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))\cdot(\boldsymbol{x}-\boldsymbol{a})t\,\mathrm{d}t\\ &= \boldsymbol{g}(\boldsymbol{x}),\end{aligned}$$

where we integrated by parts with respect to t, used that $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a})) = \nabla\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))\cdot(\boldsymbol{x}-\boldsymbol{a})$, and that $\nabla\boldsymbol{g} = \nabla\boldsymbol{g}^{\mathsf{T}}$ by assumption.
(ii) Since $\nabla\times(\boldsymbol{\phi}\times\boldsymbol{\psi}) = (\nabla\cdot\boldsymbol{\psi})\boldsymbol{\phi} - (\boldsymbol{\phi}\cdot\nabla)\boldsymbol{\psi} - (\nabla\cdot\boldsymbol{\phi})\boldsymbol{\psi} + (\boldsymbol{\psi}\cdot\nabla)\boldsymbol{\phi}$, we have

$$\begin{aligned}\nabla\times\boldsymbol{P}^{\mathrm{c}}(\boldsymbol{g})(\boldsymbol{x}) &= \int_0^1 \nabla\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))\cdot(\boldsymbol{x}-\boldsymbol{a})t^2\,\mathrm{d}t - 2\int_0^1 \boldsymbol{g}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))t\,\mathrm{d}t\\ &= \boldsymbol{g}(\boldsymbol{x}) + \int_0^1 \nabla\boldsymbol{g}^{\mathsf{T}}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))\cdot(\boldsymbol{x}-\boldsymbol{a})t^2\,\mathrm{d}t - \int_0^1 \tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))t^2\,\mathrm{d}t\\ &= \boldsymbol{g}(\boldsymbol{x}),\end{aligned}$$

where we used that $\nabla\cdot\boldsymbol{g} = 0$, $\nabla\cdot(\boldsymbol{x}-\boldsymbol{a}) = 3$, and $(\boldsymbol{\psi}\cdot\nabla)(\boldsymbol{x}-\boldsymbol{a}) = \boldsymbol{\psi}$.
(iii) We have

$$\begin{aligned}\nabla\cdot\boldsymbol{P}^{\mathrm{d}}(f)(\boldsymbol{x}) &= d\int_0^1 f(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))t^{d-1}\,\mathrm{d}t + (\boldsymbol{x}-\boldsymbol{a})\cdot\int_0^1 \nabla f(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))t^d\,\mathrm{d}t\\ &= f(\boldsymbol{x}) - \int_0^1 \tfrac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))t^d\,\mathrm{d}t + (\boldsymbol{x}-\boldsymbol{a})\cdot\int_0^1 \nabla f(\boldsymbol{a}+t(\boldsymbol{x}-\boldsymbol{a}))t^d\,\mathrm{d}t\\ &= f(\boldsymbol{x}).\end{aligned}$$

Exercise 16.5 (Koszul operator). (i) Recall that $\nabla\boldsymbol{v}$ has components $(\nabla\boldsymbol{v})_{ij} = \partial_j v_i$ for all $i,j\in\{1{:}d\}$. We have $\nabla(\boldsymbol{x}\cdot\boldsymbol{v}) = \boldsymbol{v} + (\nabla\boldsymbol{v})^{\mathsf{T}}\boldsymbol{x}$ and $\boldsymbol{x}\times(\nabla\times\boldsymbol{v}) = (\nabla\boldsymbol{v})^{\mathsf{T}}\boldsymbol{x} - (\nabla\boldsymbol{v})\boldsymbol{x}$. Thus, $\nabla(\boldsymbol{x}\cdot\boldsymbol{v}) - \boldsymbol{x}\times(\nabla\times\boldsymbol{v}) = \boldsymbol{v} + (\nabla\boldsymbol{v})\boldsymbol{x}$, and applying Euler's identity to each component of $\boldsymbol{v}$, we infer that $(\nabla\boldsymbol{v})\boldsymbol{x} = (\boldsymbol{x}\cdot\nabla)\boldsymbol{v} = k\boldsymbol{v}$. In conclusion, $\nabla(\boldsymbol{x}\cdot\boldsymbol{v}) - \boldsymbol{x}\times(\nabla\times\boldsymbol{v}) = (k+1)\boldsymbol{v}$. Let us now consider the second identity. We have $\nabla\times(\boldsymbol{x}\times\boldsymbol{v}) = \boldsymbol{x}(\nabla\cdot\boldsymbol{v}) - (\boldsymbol{x}\cdot\nabla)\boldsymbol{v} + \boldsymbol{v} - (\nabla\cdot\boldsymbol{x})\boldsymbol{v} = \boldsymbol{x}(\nabla\cdot\boldsymbol{v}) - (\boldsymbol{x}\cdot\nabla)\boldsymbol{v} - 2\boldsymbol{v}$. Hence, we have $-\nabla\times(\boldsymbol{x}\times\boldsymbol{v}) + \boldsymbol{x}(\nabla\cdot\boldsymbol{v}) = (\boldsymbol{x}\cdot\nabla)\boldsymbol{v} + 2\boldsymbol{v} = (k+2)\boldsymbol{v}$, where we used again Euler's identity to conclude.
(ii) Let us prove that $\mathbb{P}^{\mathrm{H}}_{k,d} = \nabla\mathbb{P}^{\mathrm{H}}_{k+1,d}\oplus\boldsymbol{x}\times\mathbb{P}^{\mathrm{H}}_{k-1,d}$. Let $\boldsymbol{v}\in\mathbb{P}^{\mathrm{H}}_{k,d}$. Then $q := \frac{1}{k+1}\boldsymbol{x}\cdot\boldsymbol{v}\in\mathbb{P}^{\mathrm{H}}_{k+1,d}$ and $\boldsymbol{w} := -\frac{1}{k+1}\nabla\times\boldsymbol{v}\in\mathbb{P}^{\mathrm{H}}_{k-1,d}$. The above identity implies that

$$\nabla q + \boldsymbol{x}\times\boldsymbol{w} = \frac{1}{k+1}\left(\nabla(\boldsymbol{x}\cdot\boldsymbol{v}) - \boldsymbol{x}\times\nabla\times\boldsymbol{v}\right) = \boldsymbol{v}.$$

This proves that $\mathbb{P}^{\mathrm{H}}_{k,d}\subset\nabla\mathbb{P}^{\mathrm{H}}_{k+1,d} + (\boldsymbol{x}\times\mathbb{P}^{\mathrm{H}}_{k-1,d})$, and the other inclusion is evident. Moreover, the sum is direct. Indeed, if $\boldsymbol{v}\in\mathbb{P}^{\mathrm{H}}_{k,d}$ is s.t. $\boldsymbol{v} = \nabla q = \boldsymbol{x}\times\boldsymbol{w}$ for some $q\in\mathbb{P}^{\mathrm{H}}_{k+1,d}$ and some $\boldsymbol{w}\in\mathbb{P}^{\mathrm{H}}_{k-1,d}$, then $\nabla\times\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{x}\cdot\boldsymbol{v} = 0$, so that the above identity implies that $(k+1)\boldsymbol{v} = \nabla(\boldsymbol{x}\cdot\boldsymbol{v}) - \boldsymbol{x}\times(\nabla\times\boldsymbol{v}) = \boldsymbol{0}-\boldsymbol{0} = \boldsymbol{0}$, i.e., $\boldsymbol{v} = \boldsymbol{0}$. This establishes that $\mathbb{P}^{\mathrm{H}}_{k,d} = \nabla\mathbb{P}^{\mathrm{H}}_{k+1,d}\oplus(\boldsymbol{x}\times\mathbb{P}^{\mathrm{H}}_{k-1,d})$, and by decomposing polynomials into homogeneous components, we conclude that $\mathbb{P}_{k,d} = \nabla\mathbb{P}_{k+1,d}\oplus(\boldsymbol{x}\times\mathbb{P}_{k-1,d})$. Finally, the proof that $\mathbb{P}_{k,d} = \nabla\times\mathbb{P}_{k+1,d}\oplus(\boldsymbol{x}\mathbb{P}_{k-1,d})$ is similar.

Exercise 16.6 ($\nabla\cdot\mathbb{RT}_{k,d}$ and $\nabla\times\mathbb{N}_{k,3}$). (i) Let $q\in\mathbb{P}_{k,d}$ be such that $\nabla\cdot(\boldsymbol{x}q)=0$. Writing $q:=\sum_{l\in\{0:k\}}q_l^{\mathrm{H}}$ with $q_l^{\mathrm{H}}\in\mathbb{P}_{l,d}^{\mathrm{H}}$, we infer using Lemma 14.3 that $0=\nabla\cdot(\boldsymbol{x}q)=\sum_{l\in\{0:k\}}(l+d)q_l^{\mathrm{H}}$, so that all the homogeneous polynomials q_l^{H} vanish. This shows that $\nabla\cdot:\boldsymbol{x}\mathbb{P}_{k,d}\to\mathbb{P}_{k,d}$ is injective. Since $\dim(\boldsymbol{x}\mathbb{P}_{k,d})=\dim(\mathbb{P}_{k,d})$, we infer that $\nabla\cdot:\boldsymbol{x}\mathbb{P}_{k,d}\to\mathbb{P}_{k,d}$ has full rank. The surjectivity of $\nabla\cdot:\mathbb{RT}_{k,d}\to\mathbb{P}_{k,d}$ follows from $\boldsymbol{x}\mathbb{P}_{k,d}\subset\mathbb{RT}_{k,d}$.
(ii) Using the rank nullity theorem, we infer that $\dim(\mathbb{RT}_{k,d}^{\mathrm{div}=0})=\dim(\mathbb{RT}_{k,d})-\dim(\mathbb{P}_{k,d})$. Using Lemma 14.6, we obtain for $d=2$ that $\dim(\mathbb{RT}_{k,2}^{\mathrm{div}=0})=(k+1)(k+3)-\frac{1}{2}(k+1)(k+2)=\frac{1}{2}(k+1)(k+4)$, and for $d=3$,

$$\begin{aligned}\dim(\mathbb{RT}_{k,3}^{\mathrm{div}=0})&=\frac{1}{2}(k+1)(k+2)(k+4)-\frac{1}{6}(k+1)(k+2)(k+3)\\&=\frac{1}{6}(k+1)(k+2)(2k+9).\end{aligned}$$

(iii) The identity $\mathbb{RT}_{k,3}^{\mathrm{div}=0}=\nabla\times\mathbb{P}_{k+1,3}$ follows from the hint since $\mathbb{RT}_{k,3}^{\mathrm{div}=0}\subset\mathbb{P}_{k,3}$ by Lemma 14.9. We can now compute in a different way the dimension of $\dim(\mathbb{RT}_{k,3}^{\mathrm{div}=0})$. This gives

$$\begin{aligned}\dim(\mathbb{RT}_{k,3}^{\mathrm{div}=0})=\dim(\mathbb{P}_{k+1,3}^{\mathrm{c}})^\perp&=3\binom{k+4}{3}-\binom{k+5}{3}+1\\&=\frac{1}{6}(k+1)(k+2)(2k+9).\end{aligned}$$

(iv) We have $\nabla\times:\mathbb{N}_{k,3}\to\mathbb{RT}_{k,3}^{\mathrm{div}=0}$ since $\nabla\times\boldsymbol{v}\in\mathbb{P}_{k,3}\subset\mathbb{RT}_{k,3}$ (see Lemma 15.10) and $\nabla\cdot(\nabla\times\boldsymbol{v})=0$. Moreover, we have already shown that $\nabla\times\boldsymbol{v}=\boldsymbol{0}$ implies that $\boldsymbol{v}\in\nabla\mathbb{P}_{k+1,3}$. Hence $\nabla\times:\mathbb{N}_{k,3}/\nabla\mathbb{P}_{k+1,3}\to\mathbb{RT}_{k,3}^{\mathrm{div}=0}$ is injective. Now, $\dim(\nabla\mathbb{P}_{k+1,3})=\dim(\mathbb{P}_{k+1,3})-1$, so that

$$\begin{aligned}\dim(\mathbb{N}_{k,3}/\nabla\mathbb{P}_{k+1,3})&=\dim(\mathbb{N}_{k,3})-\dim(\mathbb{P}_{k+1,3})+1\\&=\frac{1}{2}(k+1)(k+3)(k+4)-\frac{1}{6}(k+2)(k+3)(k+4)+1\\&=\frac{1}{6}(k+1)(k+2)(2k+9)=\dim(\mathbb{RT}_{k,3}^{\mathrm{div}=0}),\end{aligned}$$

owing to Step (iii). The rank nullity theorem implies that $\nabla\times:\mathbb{N}_{k,3}\to\mathbb{RT}_{k,3}^{\mathrm{div}=0}$ is surjective.

Exercise 16.7 ($\nabla\mathbb{P}_{k+1,d}$ and $\nabla\times\mathbb{P}_{k+1,3}$). (i) Let $\nabla:\mathbb{P}_{k+1,d}\longrightarrow\mathbb{P}_{k,d}$. The rank nullity theorem says that $\dim(\ker\nabla)+\dim(\mathrm{im}\,\nabla)=\dim(\mathbb{P}_{k+1,d})=\binom{k+1+d}{d}$. Since $\dim(\ker\nabla)=1$, we have $\dim(\mathbb{P}_{k,d}^{\mathrm{c}})=\dim(\mathrm{im}\,\nabla)=\binom{k+1+d}{d}-1$.
(ii) We have $\mathbb{P}_{k,3}^{\mathrm{d}}=\mathrm{im}(\nabla\times)$. The first equality follows from

$$\begin{aligned}\dim(\mathbb{P}_{k,3}^{\mathrm{d}})&=\dim(\mathbb{P}_{k+1,3})-\dim(\ker\nabla\times)\\&=\dim(\mathbb{P}_{k+1,3})-\dim(\mathrm{im}\,\nabla)\\&=\dim(\mathbb{P}_{k+1,3})-\dim(\mathbb{P}_{k+2,3})+\dim(\ker\nabla)\\&=\dim(\mathbb{P}_{k+1,3})-\dim(\mathbb{P}_{k+2,3})+1=3\binom{k+4}{3}-\binom{k+5}{3}-1,\end{aligned}$$

where we used the rank nullity theorem, that $\ker(\nabla\times)=\mathrm{im}(\nabla)$, the rank nullity theorem again, and that $\ker(\nabla)$ is composed of constant functions. The second equality follows from

$$\begin{aligned}\dim(\mathbb{P}_{k,3}^{\mathrm{d}})&=\dim(\mathrm{im}\,\nabla\times)=\dim(\ker\nabla\cdot)\\&=\dim(\mathbb{P}_{k,3})-\dim(\mathrm{im}\,\nabla\cdot)=3\binom{k+3}{3}-\binom{k+2}{3},\end{aligned}$$

where we used that $\mathrm{im}(\nabla\times)=\ker(\nabla\cdot)$, the rank nullity theorem, and the surjectivity of $\nabla\cdot$.

Chapter 17

Local interpolation in $\boldsymbol{H}(\mathrm{div})$ and $\boldsymbol{H}(\mathrm{curl})$ (II)

Exercises

Exercise 17.1 (Lifting). Let $D := (0,1)^2$. Let $\boldsymbol{x} := (x_1, x_2)^\mathsf{T}$ and consider the function $\phi(\boldsymbol{x}) := \frac{x_1}{\sqrt{x_1^2+x_2^2}}$. (i) Compute $\lim_{x_1\downarrow 0}\phi(\boldsymbol{x})$ and $\lim_{x_2\downarrow 0}\phi(\boldsymbol{x})$. (ii) Without invoking a trace argument, prove directly that $\phi \notin H^1(D)$. (iii) Construct a function $\psi \in C^\infty(D;[0,1])$ s.t. $\lim_{x_1\downarrow 0}\psi(\boldsymbol{x}) = 0$, $\lim_{x_2\uparrow 1}\psi(\boldsymbol{x}) = 0$, $\lim_{x_1\uparrow 1}\psi(\boldsymbol{x}) = 0$, and $\lim_{x_2\downarrow 0}\psi(\boldsymbol{x}) = 1$.

Exercise 17.2 (Extended face dofs for $\mathbb{RT}_{k,d}$). (i) Let $\epsilon_{K,F} := \boldsymbol{n}_F\cdot\boldsymbol{n}_{K|F}$, $\epsilon_{\widehat{K},\widehat{F}} := \boldsymbol{n}_{\widehat{F}}\cdot\boldsymbol{n}_{\widehat{K}|\widehat{F}}$, and $\epsilon_K := \det(\mathbb{J}_K)/|\det(\mathbb{J}_K)|$. Prove that $\epsilon_{K,F} = \epsilon_{\widehat{K},\widehat{F}}\epsilon_K$. (ii) Prove (17.17). (*Hint*: show that $L_F^K(\zeta_m\circ\boldsymbol{T}_{K,F}^{-1}) = L_{\widehat{F}}^{\widehat{K}}(\zeta_m\circ\boldsymbol{T}_{\widehat{F}}^{-1})\circ\boldsymbol{T}_K^{-1}$ and use (9.8a).)

Exercise 17.3 ($\mathcal{I}_K^{\mathrm{c}}$). (i) Let $r > \frac{1}{2}$ and $p \in (2, \frac{4}{3-2r}]$. Prove the stability estimate $\|\mathcal{I}_K^{\mathrm{c}}(\boldsymbol{v})\|_{\boldsymbol{L}^2(K)} \le c\left(\|\boldsymbol{v}\|_{\boldsymbol{L}^2(K)} + h_K^r|\boldsymbol{v}|_{\boldsymbol{H}^r(K)} + h_K^{1+3(\frac{1}{2}-\frac{1}{p})}\|\nabla\times\boldsymbol{v}\|_{\boldsymbol{L}^p(K)}\right)$ for all $\boldsymbol{v}\in\boldsymbol{V}^{\mathrm{c}}(K)$. (*Hint*: use the trace theorem (Theorem 3.10), the Sobolev embedding theorem (Theorem 2.31), and reason as in the proof of Theorem 17.5.) (ii) Prove Theorem 17.11. (*Hint*: proceed as in the proof of Theorem 17.5.)

Exercise 17.4 (Extended edge dofs for $\mathbb{N}_{k,d}$). Use the notation from Remark 17.10. (i) Let $\boldsymbol{w}\in\boldsymbol{C}^1(K)$ be a smooth function. Prove that $\epsilon_{K,F,E} = \epsilon_K\epsilon_{\widehat{K},\widehat{F},\widehat{E}}$ where $\epsilon_K := \det(\mathbb{J}_K)/|\det(\mathbb{J}_K)|$. (*Hint*: apply the Kelvin–Stokes formula (16.15) to the shape function of the lowest-order Nédélec element associated with E). (ii) Prove (17.28). (*Hint*: proceed as in Exercise 17.2(ii) and use (9.8b).)

Solution to exercises

Exercise 17.1 (Lifting). (i) We have

$$\lim_{x_1\downarrow 0}\phi(\boldsymbol{x}) = 0 \quad\text{and}\quad \lim_{x_2\downarrow 0}\phi(\boldsymbol{x}) = 1.$$

(ii) Using polar coordinates, we have $\phi(\boldsymbol{x}) = \cos(\theta)$. Hence, $\nabla\phi(\boldsymbol{x}) = (0, -\frac{1}{r}\sin(\theta))^\mathsf{T}$. This implies that

$$|\phi|^2_{H^1(D)} \geq \int_0^1 \int_0^{\frac{\pi}{2}} \frac{1}{r^2}\sin(\theta)^2 r \,\mathrm{d}r\,\mathrm{d}\theta = \frac{\pi}{4}\int_0^1 \frac{1}{r}\,\mathrm{d}r = \infty.$$

This proves that $\phi \notin H^1(D)$.
(iii) The following function satisfies the requirements in the question:

$$\psi(\boldsymbol{x}) = \frac{x_1}{\sqrt{x_1^2+x_2^2}}\frac{1-x_1}{\sqrt{(1-x_1)^2+x_2^2}}(1-x_2).$$

One can verify that $\psi \in H^s(D)$ for all $s \in [0,1)$.

Exercise 17.2 (Extended face dofs for $\mathbb{RT}_{k,d}$). (i) Since the orientation of the mesh $\mathcal{T}_h$ is generation-compatible according to Definition 10.3, the unit normal vectors $\boldsymbol{n}_F$ and $\boldsymbol{n}_{\widehat{F}}$ are connected by $\boldsymbol{n}_F = \boldsymbol{\Phi}^{\mathrm{d}}_K(\widehat{\boldsymbol{n}}_{\widehat{F}})$, and recalling the definition (9.14a) of $\boldsymbol{\Phi}^{\mathrm{d}}_K$ leads to

$$\boldsymbol{n}_F = \epsilon_K \frac{1}{\|\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{F}}\|_{\ell^2}}\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{F}}.$$

Moreover, Lemma 9.11 implies that

$$\boldsymbol{n}_{K|F} = \frac{1}{\|\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}}\|_{\ell^2}}\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}}.$$

Hence, we have

$$\begin{aligned}\frac{1}{\|\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}}\|_{\ell^2}}\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}} = \boldsymbol{n}_{K|F} &= \epsilon_{K,F}\boldsymbol{n}_F\\ &= \epsilon_{K,F}\epsilon_K\frac{1}{\|\mathbb{J}_K\widehat{\boldsymbol{n}}_{\widehat{F}}\|_{\ell^2}}\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{F}}\\ &= \epsilon_{K,F}\epsilon_K\epsilon_{\widehat{K},\widehat{F}}\frac{1}{\|\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}}\|_{\ell^2}}\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}}.\end{aligned}$$

This proves that $\epsilon_{K,F} = \epsilon_K\epsilon_{\widehat{K},\widehat{F}}$.
(ii) By definition, we have

$$\begin{aligned}\sigma^{\mathrm{f}}_{F,m}(\boldsymbol{v}) &= \widehat{\sigma}^{\mathrm{f}}_{\widehat{F},m}(\boldsymbol{\psi}^{\mathrm{d}}_K(\boldsymbol{v}))\\ &= \epsilon_{\widehat{K},\widehat{F}}\int_{\widehat{K}}\Big(\boldsymbol{\psi}^{\mathrm{d}}_K(\boldsymbol{v})\cdot\nabla L^{\widehat{K}}_{\widehat{F}}(\zeta_m\circ\boldsymbol{T}^{-1}_{\widehat{F}}) + (\nabla\cdot\boldsymbol{\psi}^{\mathrm{d}}_K(\boldsymbol{v}))L^{\widehat{K}}_{\widehat{F}}(\zeta_m\circ\boldsymbol{T}^{-1}_{\widehat{F}})\Big)\,\mathrm{d}\widehat{x}.\end{aligned}$$

Now, for all $\boldsymbol{x} \in F = \boldsymbol{T}_K(\widehat{F})$, we use the definition of L^K_F stated in (17.9). This gives

$$\begin{aligned}L^K_F(\zeta_m\circ\boldsymbol{T}^{-1}_{K,F})(\boldsymbol{x}) &= L^K_F(\zeta_m\circ\boldsymbol{T}^{-1}_{\widehat{F}}\circ\boldsymbol{T}^{-1}_{K|\widehat{F}})(\boldsymbol{x})\\ &= L^{\widehat{K}}_{\widehat{F}}(\zeta_m\circ\boldsymbol{T}^{-1}_{\widehat{F}}\circ\boldsymbol{T}^{-1}_{K|\widehat{F}}\circ\boldsymbol{T}_{K|\widehat{F}})(\boldsymbol{T}^{-1}_K(\boldsymbol{x}))\\ &= L^{\widehat{K}}_{\widehat{F}}(\zeta_m\circ\boldsymbol{T}^{-1}_{\widehat{F}})(\boldsymbol{T}^{-1}_K(\boldsymbol{x})),\end{aligned}$$

where we used that $\boldsymbol{T}_{K,F}^{-1} = \boldsymbol{T}_{\widehat{F}}^{-1} \circ \boldsymbol{T}_{K|\widehat{F}}^{-1}$. Owing to (9.8a) and since $\mathbb{J}_K$ is constant, we infer that

$$\begin{aligned}\nabla(L_{\widehat{F}}^{\widehat{K}}(\zeta_m \circ \boldsymbol{T}_{\widehat{F}}^{-1}))(\widehat{\boldsymbol{x}}) &= \nabla(L_F^K(\zeta_m \circ \boldsymbol{T}_{K,F}^{-1}) \circ \boldsymbol{T}_K)(\widehat{\boldsymbol{x}}) \\ &= \mathbb{J}_K^{\mathsf{T}} \nabla(L_F^K(\zeta_m \circ \boldsymbol{T}_{K,F}^{-1}))(\boldsymbol{T}_K(\widehat{\boldsymbol{x}})).\end{aligned}$$

This, in turn, implies that

$$\begin{aligned}\sigma_{F,m}^{\mathrm{f}}(\boldsymbol{v}) &= \epsilon_{\widehat{K},\widehat{F}} \int_{\widehat{K}} \Big(\boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{v}) \cdot \nabla(L_{\widehat{F}}^{\widehat{K}}(\zeta_m \circ \boldsymbol{T}_{\widehat{F}}^{-1})) + (\nabla \cdot \boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{v})) L_{\widehat{F}}^{\widehat{K}}(\zeta_m \circ \boldsymbol{T}_{\widehat{F}}^{-1}) \Big) \,\mathrm{d}\widehat{x} \\ &= \epsilon_{\widehat{K},\widehat{F}} \int_{\widehat{K}} \Big(\det(\mathbb{J}_K) \mathbb{J}_K^{-1} \boldsymbol{v} \cdot \mathbb{J}_K^{\mathsf{T}} \nabla(L_F^K(\zeta_m \circ \boldsymbol{T}_{K,F}^{-1})) \\ &\qquad + \det(\mathbb{J}_K)(\nabla \cdot \boldsymbol{v}) L_F^K(\zeta_m \circ \boldsymbol{T}_{K,F}^{-1}) \Big)(\boldsymbol{T}_K(\widehat{\boldsymbol{x}})) \,\mathrm{d}\widehat{x} \\ &= \epsilon_{\widehat{K},\widehat{F}} \epsilon_K \int_K \Big(\boldsymbol{v} \cdot \nabla(L_F^K(\zeta_m \circ \boldsymbol{T}_{K,F}^{-1})) + (\nabla \cdot \boldsymbol{v}) L_F^K(\zeta_m \circ \boldsymbol{T}_{K,F}^{-1}) \Big)(\boldsymbol{x}) \,\mathrm{d}x,\end{aligned}$$

with $\epsilon_K := \det(\mathbb{J}_K)/|\det(\mathbb{J}_K)|$. We conclude using the identity $\epsilon_{K,F} = \epsilon_K \epsilon_{\widehat{K},\widehat{F}}$ from Step (i).

Exercise 17.3 ($\mathcal{I}_K^{\mathrm{c}}$). (i) Using Proposition 12.5, $\mathbb{A}_K^{\mathrm{c}} := \mathbb{J}_K^{\mathsf{T}}$, and the regularity of the mesh sequence, we infer that

$$\|\mathcal{I}_K^{\mathrm{c}}(\boldsymbol{v})\|_{\boldsymbol{L}^2(K)} \le c\, h_K^{\frac{3}{2}-1} \max_{i \in \mathcal{N}} |\sigma_{K,i}(\boldsymbol{v})|.$$

Proposition 17.9 leads to

$$\|\mathcal{I}_K^{\mathrm{c}}(\boldsymbol{v})\|_{\boldsymbol{L}^2(K)} \le c\, h_K^{\frac{3}{2}} \Big(h_K^{-\frac{3}{p}} \|\boldsymbol{v}\|_{\boldsymbol{L}^p(K)} + h_K^{1-\frac{3}{p}} \|\nabla \times \boldsymbol{v}\|_{\boldsymbol{L}^p(K)} + h_K^{-\frac{2}{p}} \|\boldsymbol{v} \times \boldsymbol{n}_K\|_{\boldsymbol{L}^p(\partial K)} \Big).$$

Since $\boldsymbol{v} \in \boldsymbol{H}^r(K)$, $r > \frac{1}{2}$, the trace theorem (Theorem 3.10) implies that $\boldsymbol{v} \times \boldsymbol{n}_K \in \boldsymbol{L}^p(\partial K)$ since $p \in (2, \frac{4}{3-2r}]$. The Sobolev embedding theorem (Theorem 2.31) implies that $\boldsymbol{v} \in \boldsymbol{L}^q(K)$ with $q := \frac{6}{3-2r} > p$. Reasoning as in the proof of Theorem 17.5, we infer that

$$h_K^{-\frac{3}{p}} \|\boldsymbol{v}\|_{\boldsymbol{L}^p(K)} + h_K^{-\frac{2}{p}} \|\boldsymbol{v} \times \boldsymbol{n}_K\|_{\boldsymbol{L}^p(\partial K)} \le c \big(h_K^{-\frac{3}{2}} \|\boldsymbol{v}\|_{\boldsymbol{L}^2(K)} + h_K^{r-\frac{3}{2}} |\boldsymbol{v}|_{\boldsymbol{H}^r(K)} \big).$$

This leads to

$$\|\mathcal{I}_K^{\mathrm{c}}(\boldsymbol{v})\|_{\boldsymbol{L}^2(K)} \le c \left(\|\boldsymbol{v}\|_{\boldsymbol{L}^2(K)} + h_K^r |\boldsymbol{v}|_{\boldsymbol{H}^r(K)} + h_K^{1+3(\frac{1}{2}-\frac{1}{p})} \|\nabla \times \boldsymbol{v}\|_{\boldsymbol{L}^p(K)} \right),$$

which is the expected stability bound.
(ii) We can now conclude by proceeding as in the proof of Theorem 17.5. We combine the stability bound from Step (i) with the fact that $\mathbb{P}_{0,d}$ is pointwise invariant under $\mathcal{I}_K^{\mathrm{c}}$, the fractional Poincaré–Steklov inequality from Lemma 12.12, and that $|\boldsymbol{v} - \boldsymbol{q}|_{\boldsymbol{H}^r(K)} = |\boldsymbol{v}|_{\boldsymbol{H}^r(K)}$ and $\nabla \times (\boldsymbol{v} - \boldsymbol{q}) = \nabla \times \boldsymbol{v}$ for all $\boldsymbol{q} \in \mathbb{P}_{0,d}$.

Exercise 17.4 (Extended edge dofs). (i) Let $\boldsymbol{\theta}_E^{\mathrm{e}}$ be the shape function of the lowest-order

Nédélec element associated with the edge E. We have

$$\begin{aligned}
\epsilon_{K,F,E} &= \epsilon_{K,F,E}\sigma_E^{\mathrm{e}}(\boldsymbol{\theta}_E^{\mathrm{e}}) = \epsilon_{K,F,E}\frac{1}{|E|}\int_E(\boldsymbol{\theta}_E^{\mathrm{e}}\cdot\boldsymbol{t}_E)\,\mathrm{d}l = \epsilon_{K,F,E}\int_E(\boldsymbol{\theta}_E^{\mathrm{e}}\cdot\boldsymbol{\tau}_E)\,\mathrm{d}l\\
&= \int_E(\boldsymbol{\theta}_E^{\mathrm{e}}\cdot\boldsymbol{\tau}_{K,F|E})\,\mathrm{d}l = \int_{\partial F}(\boldsymbol{\theta}_E^{\mathrm{e}}\cdot\boldsymbol{\tau}_{K,F})\,\mathrm{d}l\\
&= \int_F(\nabla\times\boldsymbol{\theta}_E^{\mathrm{e}})\cdot\boldsymbol{n}_{K|F}\,\mathrm{d}s = \int_F(\mathbb{J}_K^{-1}(\nabla\times\boldsymbol{\theta}_E^{\mathrm{e}}))\cdot(\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F})\,\mathrm{d}s\\
&= \int_F\det(\mathbb{J}_K^{-1})\big((\nabla\times\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{\theta}_E^{\mathrm{e}}))\cdot\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}}\big)\circ\boldsymbol{T}_{K[F}^{-1}\|\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F}\|_{\ell^2}\,\mathrm{d}s\\
&= \epsilon_K\int_{\widehat{F}}(\nabla\times\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{\theta}_E^{\mathrm{e}}))\cdot\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}}\,\mathrm{d}\widehat{s},
\end{aligned}$$

where we used the definition of the $\mathbb{N}_{0,d}$ dofs and of the tangent vectors $\boldsymbol{t}_E$ and $\boldsymbol{\tau}_E$ in the first line, the definition of $\epsilon_{K,F,E}$ and the fact that $\boldsymbol{\theta}_E^{\mathrm{e}}$ has zero tangential component on $\partial F\backslash E$ in the second line, the Kelvin–Stokes formula (16.15) and an elementary manipulation in the third line, the identity (9.8b) and the fact that $\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F} = \|\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F}\|_{\ell^2}\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}}\circ\boldsymbol{T}_{K|F}^{-1}$ in the fourth line, and the transformation of surface measures in the fifth line. Since $\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{\theta}_E^{\mathrm{e}})$ is the reference shape function of the lowest-order Nédélec element associated with the edge $\widehat{E}$ s.t. $\boldsymbol{T}_K(\widehat{E}) = E$, we conclude using the same arguments as above that

$$\epsilon_{\widehat{K},\widehat{F},\widehat{E}} = \epsilon_{\widehat{K},\widehat{F},\widehat{E}}\widehat{\sigma}_E^{\mathrm{e}}(\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{\theta}_E^{\mathrm{e}})) = \int_{\widehat{F}}(\nabla\times\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{\theta}_E^{\mathrm{e}}))\cdot\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}}\,\mathrm{d}\widehat{s},$$

and putting everything together yields the expected identity.
(ii) Let $\boldsymbol{v}\in\boldsymbol{V}^{\mathrm{c}}(K)$. By definition, we have $\sigma_{E,m}^{\mathrm{e}} = \widehat{\sigma}_{\widehat{E},m}^{\mathrm{e}}(\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{v})) = \epsilon_{\widehat{K},\widehat{F},\widehat{E}}(\mathfrak{T}_1+\mathfrak{T}_2)$ with

$$\begin{aligned}
\mathfrak{T}_1 &:= \int_{\widehat{K}}(\nabla\times\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{v}))\cdot\nabla L_{\widehat{E}}^{\widehat{K}}(\mu_m\circ\boldsymbol{T}_{\widehat{E}}^{-1})\,\mathrm{d}\widehat{x},\\
\mathfrak{T}_2 &:= \int_{\widehat{F}}(\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{v})\times\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}})\cdot\nabla L_{\widehat{E}}^{\widehat{F}}(\mu_m\circ\boldsymbol{T}_{\widehat{E}}^{-1})\,\mathrm{d}\widehat{s}.
\end{aligned}$$

Proceeding as in Step (ii) of Exercise 17.2, we infer that $L_{\widehat{E}}^{\widehat{K}}(\mu_m\circ\boldsymbol{T}_{\widehat{E}}^{-1}) = L_E^K(\mu_m\circ\boldsymbol{T}_{K,E}^{-1})\circ\boldsymbol{T}_K$, so that invoking (9.8a) we obtain

$$\nabla L_{\widehat{E}}^{\widehat{K}}(\mu_m\circ\boldsymbol{T}_{\widehat{E}}^{-1}) = \mathbb{J}_K^{\mathsf{T}}(\nabla L_E^K(\mu_m\circ\boldsymbol{T}_{K,E}^{-1}))\circ\boldsymbol{T}_K.$$

Invoking (9.8b), we infer that

$$\begin{aligned}
\mathfrak{T}_1 &= \int_{\widehat{K}}\det(\mathbb{J}_K)\big(\mathbb{J}_K^{-1}(\nabla\times\boldsymbol{v})\cdot\mathbb{J}_K^{\mathsf{T}}(\nabla L_E^K(\mu_m\circ\boldsymbol{T}_{K,E}^{-1}))\big)\circ\boldsymbol{T}_K\,\mathrm{d}\widehat{x}\\
&= \epsilon_K\int_K(\nabla\times\boldsymbol{v})\cdot\nabla L_E^K(\mu_m\circ\boldsymbol{T}_{K,E}^{-1})\,\mathrm{d}x.
\end{aligned}$$

Similarly, we have $\nabla L_{\widehat{E}}^{\widehat{F}}(\mu_m\circ\boldsymbol{T}_{\widehat{E}}^{-1}) = \mathbb{J}_K^{\mathsf{T}}(\nabla L_E^F(\mu_m\circ\boldsymbol{T}_{K,E}^{-1}))\circ\boldsymbol{T}_K$, and

$$\begin{aligned}
\boldsymbol{\psi}_K^{\mathrm{c}}(\boldsymbol{v})\times\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}} &= \frac{1}{\|\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F}\|_{\ell^2}}((\mathbb{J}_K^{\mathsf{T}}\boldsymbol{v})\times(\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F}))\circ\boldsymbol{T}_{K|\widehat{F}}\\
&= \det(\mathbb{J}_K)\frac{1}{\|\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F}\|_{\ell^2}}\mathbb{J}_K^{-1}(\boldsymbol{v}\times\boldsymbol{n}_{K|F})\circ\boldsymbol{T}_{K|\widehat{F}},
\end{aligned}$$

where we used the definition of $\boldsymbol{\psi}_K^{\mathrm{c}}$, the identity (9.10) on the transformation of unit normals, and $\|(\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F})(\boldsymbol{x})\|_{\ell^2} = \|(\mathbb{J}_K^{-\mathsf{T}}\widehat{\boldsymbol{n}}_{\widehat{K}|\widehat{F}})(\widehat{\boldsymbol{x}})\|_{\ell^2}^{-1}$ in the first line, and the identity from Exercise 9.5 in the second line. Using Lemma 9.12 on the transformation of surface measures, we infer that

$$\begin{aligned}
\mathfrak{T}_2 &= \int_{\widehat{F}} \big((\mathbb{J}_K^{\mathsf{T}}\boldsymbol{v}\times\frac{\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F}}{\|\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F}\|_{\ell^2}})\cdot\mathbb{J}_K^{\mathsf{T}}(\nabla L_E^F(\mu_m\circ\boldsymbol{T}_{K,E}^{-1}))\big)\circ\boldsymbol{T}_{K|\widehat{F}}\,\mathrm{d}\widehat{s} \\
&= \int_{\widehat{F}} \det(\mathbb{J}_K)\frac{1}{\|\mathbb{J}_K^{\mathsf{T}}\boldsymbol{n}_{K|F}\|_{\ell^2}}\big(\mathbb{J}_K^{-1}(\boldsymbol{v}\times\boldsymbol{n}_{K|F})\cdot\mathbb{J}_K^{\mathsf{T}}(\nabla L_E^F(\mu_m\circ\boldsymbol{T}_{K,E}^{-1}))\big)\circ\boldsymbol{T}_{K|\widehat{F}}\,\mathrm{d}\widehat{s} \\
&= \epsilon_K\int_F(\boldsymbol{v}\times\boldsymbol{n}_{K|F})\cdot\nabla L_E^F(\mu_m\circ\boldsymbol{T}_{K,E}^{-1})\,\mathrm{d}s.
\end{aligned}$$

Putting everything together and using the identity $\epsilon_{K,F,E} = \epsilon_K\epsilon_{\widehat{K},\widehat{F},\widehat{E}}$ from Step (i) proves that (17.28) holds true.

Chapter 18

From broken to conforming spaces

Exercises

Exercise 18.1 ($\boldsymbol{H}$(div), $\boldsymbol{H}$(curl)). Prove Theorem 18.10. (*Hint*: use (4.8).)

Exercise 18.2 (Discrete Sobolev inequality). (i) Assume $d \geq 3$. Prove that $\|v_h\|_{L^\infty(K)} \leq c h_K^{1-\frac{d}{2}} \|\nabla v_h\|_{\boldsymbol{L}^2(K)}$ for all $v_h \in P_k^{\mathrm{g,b}}(\mathcal{T}_h)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$. (*Hint*: use Theorem 2.31.) (ii) Assume $d = 2$. Prove (18.15). (*Hint*: let $K \in \mathcal{T}_h$ with $h_K \leq \frac{\delta_D}{2}$, let $\boldsymbol{x} \in K$ and let $\boldsymbol{y}$ have polar coordinates (r, θ) with respect to $\boldsymbol{x}$ with $r \geq \frac{\delta_D}{2}$ and $\theta \in (0, \omega)$, use that $v_h(\boldsymbol{x}) = v_h(\boldsymbol{y}) - \int_0^r \partial_\rho v_h(\rho, \theta) \,\mathrm{d}\rho$, decompose the integral as $\int_0^r \cdot \,\mathrm{d}\rho = \int_0^{h_K} \cdot \,\mathrm{d}\rho + \int_{h_K}^r \cdot \,\mathrm{d}\rho$, and bound the two addends.)

Exercise 18.3 (Orthogonal and oblique projections). (i) Show that $\mathcal{I}_{\widehat{K}}^\sharp$ is the L^2-orthogonal projection onto $\widehat{P}$. (*Hint*: observe that $(\widehat{\rho}_i, \widehat{\theta}_j)_{L^2(\widehat{K};\mathbb{R}^q)} = |\widehat{K}|\delta_{ij}$ for all $i, j \in \mathcal{N}$.) (ii) Prove that $\mathcal{I}_K^\sharp$ is the oblique projection onto $P_K = \psi_K^{-1}(\widehat{P})$ parallel to $Q_K^\perp$ with $Q_K := \Phi_K^{-1}(\widehat{P})$. (*Hint*: use (18.17).) (iii) Show that $P_K = Q_K$ if the matrix $\mathbb{A}_K$ is unitary, i.e., $\mathbb{A}_K^\mathsf{T}\mathbb{A}_K = \mathbb{A}_K\mathbb{A}_K^\mathsf{T} = \mathbb{I}_q$.

Exercise 18.4 (Approximation on faces). Prove (18.28).

Solution to exercises

Exercise 18.1 ($\boldsymbol{H}$(div), $\boldsymbol{H}$(curl)). Let $\boldsymbol{v} \in \boldsymbol{W}^{1,p}(\mathcal{T}_h)$. Using the hint, we infer for the divergence that

$$\int_D \boldsymbol{v}\cdot\nabla\Phi \,\mathrm{d}x = \sum_{K\in\mathcal{T}_h} -\int_K \nabla\cdot(\boldsymbol{v}_{|K})\Phi \,\mathrm{d}x + \sum_{F\in\mathcal{F}_h^\circ}\int_F [\![\boldsymbol{v}\cdot\boldsymbol{n}]\!]_F \Phi \,\mathrm{d}s,$$

for all $\Phi \in C_0^\infty(D)$, and we infer for the curl that

$$\int_D \boldsymbol{v}\cdot\nabla\times\boldsymbol{\Phi} \,\mathrm{d}x = \sum_{K\in\mathcal{T}_h} \int_K \nabla\times(\boldsymbol{v}_{|K})\cdot\boldsymbol{\Phi} \,\mathrm{d}x + \sum_{F\in\mathcal{F}_h^\circ}\int_F [\![\boldsymbol{v}\times\boldsymbol{n}]\!]_F\cdot\boldsymbol{\Phi} \,\mathrm{d}s,$$

for all $\boldsymbol{\Phi} \in \boldsymbol{C}_0^\infty(D)$. The rest of the proof follows the same arguments as those presented in the proof of Theorem 18.8.

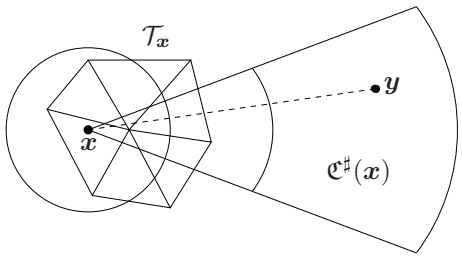
Exercise 18.2 (Discrete Sobolev inequality). (i) It suffices to apply the inverse inequality (12.3) with $p := \infty$ and $q := 2^* := \frac{2d}{d-2}$ followed by Theorem 2.31.
(ii) Let $v_h \in P_k^{\mathrm{g}}(\mathcal{T}_h)$ and let $K \in \mathcal{T}_h$ be such that $h_K \le \frac{\delta_D}{2}$. Fix $\boldsymbol{x} \in K$. Set $\mathfrak{C}^\sharp(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathfrak{C}(\boldsymbol{x}) \mid \|\boldsymbol{y}-\boldsymbol{x}\|_{\ell^2(\mathbb{R}^2)} \ge \frac{\delta_D}{2}\}$. Let $\boldsymbol{y}$ be arbitrary in $\mathfrak{C}^\sharp(\boldsymbol{x})$ with polar coordinates (r,θ) with respect to $\boldsymbol{x}$. Since $v_h(\boldsymbol{x}) = v_h(\boldsymbol{y}) - \int_0^r \partial_\rho v_h(\rho,\theta)\,\mathrm{d}\rho$, we infer that

$$|v_h(\boldsymbol{x})|^2 \le 2|v_h(\boldsymbol{y})|^2 + 2(I_1+I_2)^2,$$

with

$$I_1 := \int_0^{h_K} \partial_\rho v_h(\rho,\theta)\,\mathrm{d}\rho, \qquad I_2 := \int_{h_K}^r \partial_\rho v_h(\rho,\theta)\,\mathrm{d}\rho.$$

Concerning I_1, let $B(\boldsymbol{x},h_K)$ be the ball of center $\boldsymbol{x}$ and radius h_K and set $\mathcal{T}_{\boldsymbol{x}} := \{K' \in \mathcal{T}_h \mid K' \cap (B(\boldsymbol{x},h_K)\cap\mathfrak{C}(\boldsymbol{x})) \neq \emptyset\}$, as illustrated in the figure below.



By definition, we have $|I_1| \le h_K \max_{K'\in\mathcal{T}_{\boldsymbol{x}}} \|\nabla v_h\|_{\boldsymbol{L}^\infty(K')}$. Using the inverse inequality (12.3) with $p := \infty$ and $q := 2$ and the fact that all the mesh cells in $\mathcal{T}_{\boldsymbol{x}}$ have a size equivalent to that of K owing to Proposition 11.6, we infer that

$$|I_1| \le c\,h_K \max_{K'\in\mathcal{T}_{\boldsymbol{x}}} h_{K'}^{-1}\|\nabla v_h\|_{\boldsymbol{L}^2(K')} \le c\,\|\nabla v_h\|_{\boldsymbol{L}^2(B(\boldsymbol{x},h_K)\cap\mathfrak{C}(\boldsymbol{x}))}.$$

Concerning I_2, we employ the Cauchy–Schwarz inequality to infer that

$$|I_2|^2 = \left(\int_{h_K}^r \rho^{-\frac12}\rho^{\frac12}\partial_\rho v_h(\rho,\theta)\,\mathrm{d}\rho\right)^2 \le \ln\left(\frac{r}{h_K}\right)\int_{h_K}^r |\partial_\rho v_h(\rho,\theta)|^2\rho\,\mathrm{d}\rho,$$

and the logarithmic factor is bounded by $\ln(\frac{\delta_D}{h_K})$. We regroup the above bounds on I_1 and I_2 and integrate the inequality for all $\boldsymbol{y} \in \mathfrak{C}^\sharp(\boldsymbol{x})$ to infer that there is $c>0$ such that

$$c\,|\mathfrak{C}^\sharp(\boldsymbol{x})||v_h(\boldsymbol{x})|^2 \le \|v_h\|_{L^2(D)}^2 + |\mathfrak{C}^\sharp(\boldsymbol{x})|\|\nabla v_h\|_{\boldsymbol{L}^2(D)}^2 + \ln\left(\frac{\delta_D}{h_K}\right)\delta_D^2\|\nabla v_h\|_{\boldsymbol{L}^2(D)}^2,$$

where we bounded integrals over $\mathfrak{C}(\boldsymbol{x})$ by integrals over D and where we used that

$$\int_{\frac12\delta_D}^{\delta_D}\int_0^\omega\int_0^r |\partial_\rho v_h(\rho,\theta)|^2\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\,\mathrm{d}r \le \delta_D^2\|\nabla v_h\|_{L^2(D)}^2$$

to bound the last term on the right-hand side. The assertion follows by dividing by $|\mathfrak{C}^\sharp(\boldsymbol{x})|$ which scales like $\omega\delta_D^2$ with $\omega > 0$.

Exercise 18.3 (Orthogonal and oblique projections). (i) Since we have $\mathrm{im}(\mathcal{I}_{\widehat{K}}^\sharp) \subset \widehat{P}$, we only need to prove that $(\widehat{q}, \mathcal{I}_{\widehat{K}}^\sharp(\widehat{v}))_{L^2(\widehat{K};\mathbb{R}^q)} = (\widehat{q},\widehat{v})_{L^2(\widehat{K};\mathbb{R}^q)}$ for all $\widehat{q} \in \widehat{P}$ and all $\widehat{v} \in L^1(\widehat{K};\mathbb{R}^q)$. The definition of the dofs implies that for all $i,j \in \mathcal{N}$,

$$\delta_{ij} = \widehat{\sigma}_j(\widehat{\theta}_i) = \frac{1}{|\widehat{K}|}(\widehat{\rho}_j,\widehat{\theta}_i)_{L^2(\widehat{K};\mathbb{R}^q)}.$$

Let now $\widehat{v} \in L^1(\widehat{K};\mathbb{R}^q)$ and let $\widehat{q} \in \widehat{P}$. Since $\{\widehat{\rho}_j\}_{j\in\mathcal{N}}$ is a basis of $\widehat{P}$, we can write $\widehat{q} = \sum_{j\in\mathcal{N}} \lambda_j \widehat{\rho}_j$. We infer that

$$\begin{aligned}
(\widehat{q}, \mathcal{I}_{\widehat{K}}^\sharp(\widehat{v}))_{L^2(\widehat{K};\mathbb{R}^q)} &= \sum_{i\in\mathcal{N}} \widehat{\sigma}_i^\sharp(\widehat{v})(\widehat{q},\widehat{\theta}_i)_{L^2(\widehat{K};\mathbb{R}^q)} \\
&= \sum_{i\in\mathcal{N}} \frac{1}{|\widehat{K}|}(\widehat{\rho}_i,\widehat{v})_{L^2(\widehat{K};\mathbb{R}^q)}(\widehat{q},\widehat{\theta}_i)_{L^2(\widehat{K};\mathbb{R}^q)} \\
&= \sum_{i,j\in\mathcal{N}} \frac{1}{|\widehat{K}|}\lambda_j(\widehat{\rho}_i,\widehat{v})_{L^2(\widehat{K};\mathbb{R}^q)}(\widehat{\rho}_j,\widehat{\theta}_i)_{L^2(\widehat{K};\mathbb{R}^q)} \\
&= \sum_{j\in\mathcal{N}} \lambda_j(\widehat{\rho}_j,\widehat{v})_{L^2(\widehat{K};\mathbb{R}^q)} = (\widehat{q},\widehat{v})_{L^2(\widehat{K};\mathbb{R}^q)},
\end{aligned}$$

thereby proving the assertion.
(ii) Since $\mathrm{im}(\mathcal{I}_K^\sharp) \subset P_K$, we only need to prove that $(q, \mathcal{I}_K^\sharp(v))_{L^2(K;\mathbb{R}^q)} = (q,v)_{L^2(K;\mathbb{R}^q)}$ for all $q \in Q_K$ and all $v \in L^1(K;\mathbb{R}^q)$. Using that $\mathcal{I}_{\widehat{K}}^\sharp$ is the L^2-orthogonal projection onto $\widehat{P}$, that $\Phi_K(q) \in \widehat{P}$, and the identity (18.17) twice, we infer that

$$\begin{aligned}
(q, \mathcal{I}_K^\sharp(v))_{L^2(K;\mathbb{R}^q)} &= (q, \psi_K^{-1}(\mathcal{I}_{\widehat{K}}^\sharp(\psi_K(v))))_{L^2(K;\mathbb{R}^q)} \\
&= (\phi_K(q), \mathcal{I}_{\widehat{K}}^\sharp(\psi_K(v)))_{L^2(\widehat{K};\mathbb{R}^q)} \\
&= (\phi_K(q), \psi_K(v))_{L^2(\widehat{K};\mathbb{R}^q)} = (q,v)_{L^2(K;\mathbb{R}^q)},
\end{aligned}$$

thereby proving the assertion.
(iii) If the matrix $\mathbb{A}_K$ is unitary, we have $\mathbb{A}_K^{-\mathsf{T}} = \mathbb{A}_K$ and since $|\det(\mathbb{J}_K)| = 1$, we infer that

$$\begin{aligned}
[\, q \in Q_K \,] &\iff [\, \phi_K(q) \in \widehat{P} \,] \\
&\iff [\, |\det(\mathbb{J}_K)|\mathbb{A}_K^{-\mathsf{T}}(q \circ \boldsymbol{T}_K) \in \widehat{P} \,] \\
&\iff [\, \mathbb{A}_K(q \circ \boldsymbol{T}_K) \in \widehat{P} \,] \\
&\iff [\, \psi_K(q) \in \widehat{P} \,] \\
&\iff [\, q \in P_K \,].
\end{aligned}$$

This shows that $P_K = Q_K$ and that $\mathcal{I}_K^\sharp$ is L^2-orthogonal.

Exercise 18.4 (Approximation on faces). Assume that $k \geq 1$ and for simplicity that $q = 1$. Let $v \in W^{1+r,p}(K)$. Assume first that $r \in [1,k]$. Owing to the multiplicative trace inequality (12.16), we infer that, with $\eta := v - \mathcal{I}_K^\sharp(v)$,

$$\|\nabla\eta\|_{\boldsymbol{L}^p(F)} \leq c\left(h_K^{-\frac{1}{p}}|\eta|_{W^{1,p}(K)} + |\eta|_{W^{1,p}(K)}^{1-\frac{1}{p}}|\eta|_{W^{2,p}(K)}^{\frac{1}{p}}\right).$$

Invoking (18.25) with $m \in \{1,2\}$ (note that $m \le 1 + \lfloor r \rfloor$) shows that (18.28) holds true in this case. Let us now assume that $r \in (\frac{1}{p}, 1)$ with $p > 1$. Let $q_1 \in \psi_K^{-1}(\mathbb{P}_{1,d}) = \mathbb{P}_{1,d}$ be arbitrary. We have

$$\begin{aligned} h_K^{\frac{1}{p}} \|\nabla \eta\|_{\boldsymbol{L}^p(F)} &\le h_K^{\frac{1}{p}} \|\nabla(v - q_1)\|_{\boldsymbol{L}^p(F)} + h_K^{\frac{1}{p}} \|\nabla(\mathcal{I}_K^\sharp(v) - q_1)\|_{\boldsymbol{L}^p(F)} \\ &\le c \left(|v - q_1|_{W^{1,p}(K)} + h_K^r |v|_{W^{1+r,p}(K)} + |\mathcal{I}_K^\sharp(v) - q_1|_{W^{1,p}(K)} \right) \\ &\le c \left(|v - q_1|_{W^{1,p}(K)} + h_K^r |v|_{W^{1+r,p}(K)} + |v - \mathcal{I}_K^\sharp(v)|_{W^{1,p}(K)} \right), \end{aligned}$$

where we used the triangle inequality in the first line, the fractional trace inequality (12.17), that $q_1 \in \mathbb{P}_{1,d}$, and the discrete trace inequality (12.10) in the second line, and the triangle inequality in the third line. Invoking (12.15) (since q_1 is arbitrary in $\mathbb{P}_{1,d}$) and (18.25) with $m := 1$ leads again to (18.28).

Chapter 19

Main properties of the conforming subspaces

Exercises

Exercise 19.1 (Connectivity classes). Consider the mesh shown in Figure 19.1 and let $P_2^{\mathrm{g}}(\mathcal{T}_h)$ be the associated finite element space composed of continuous Lagrange $\mathbb{P}_2$ finite elements. Assume that the enumeration of the Lagrange nodes has been done with the increasing vertex-index technique (see (10.10)). (i) What is the domain and the codomain of `j_dof`? (ii) Identify $\texttt{j\_dof}^{-1}(8)$ and $\texttt{j\_dof}^{-1}(13)$. (iii) Identify $\mathcal{T}_6$ and $\mathcal{T}_{10}$.

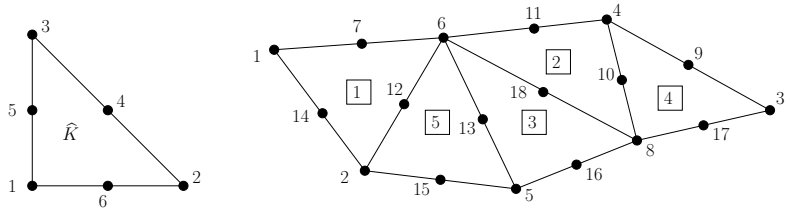


Figure 19.1: Illustration for Exercise 19.1.

Exercise 19.2 (Stiffness, mass, incidence matrices). Let $\{\lambda_n\}_{n\in\{1:N_{\mathrm{v}}\}}$ be the global shape functions in $P_1^{\mathrm{g}}(\mathcal{T}_h)$. Let $\{\boldsymbol{\theta}_m\}_{m\in\{1:N_{\mathrm{e}}\}}$ be the global shape functions in $\boldsymbol{P}_0^{\mathrm{c}}(\mathcal{T}_h)$. (i) Recall the incidence matrix $\mathcal{M}^{\mathrm{ev}} \in \mathbb{R}^{N_{\mathrm{e}}\times N_{\mathrm{v}}}$ defined in Remark 10.2. Prove that $\nabla\lambda_n = \sum_{m\in\{1:N_{\mathrm{e}}\}} \mathcal{M}^{\mathrm{ev}}_{mn}\boldsymbol{\theta}_m$ for all $n \in \{1{:}N_{\mathrm{v}}\}$. (*Hint*: compute $\sigma^{\mathrm{e}}_m(\nabla\lambda_n)$ where $\{\sigma^{\mathrm{e}}_m\}_{m\in\{1:N_{\mathrm{e}}\}}$ is the dual basis of $\{\boldsymbol{\theta}_m\}_{m\in\{1:N_{\mathrm{e}}\}}$, i.e., the associated dofs.) (ii) Let $\mathcal{A} \in \mathbb{R}^{N_{\mathrm{v}}\times N_{\mathrm{v}}}$ be the Courant stiffness matrix with entries $\mathcal{A}_{nn'} := \int_D \nabla\lambda_n\cdot\nabla\lambda_{n'}\,\mathrm{d}x$ for all $n, n' \in \{1{:}N_{\mathrm{v}}\}$, and let $\mathcal{N} \in \mathbb{R}^{N_{\mathrm{e}}\times N_{\mathrm{e}}}$ be the Nédélec mass matrix with entries $\mathcal{N}_{mm'} := \int_D \boldsymbol{\theta}_m\cdot\boldsymbol{\theta}_{m'}\,\mathrm{d}x$ for all $m, m' \in \{1{:}N_{\mathrm{e}}\}$. Prove that $\mathcal{A} = (\mathcal{M}^{\mathrm{ev}})^{\mathsf{T}}\mathcal{N}\mathcal{M}^{\mathrm{ev}}$.

Exercise 19.3 (Zero trace). (i) Show that $\varphi_a \in P^{\mathrm{x}}_{k,0}(\mathcal{T}_h)$ for all $a \in \mathcal{A}_h^\circ$. (ii) Prove Proposition 19.13.

Exercise 19.4 (Approximability in L^p). Let $p \in [1,\infty)$. Prove that $\lim_{h\downarrow 0}\inf_{v_h\in P^{\mathrm{g}}_k(\mathcal{T}_h)}\|v - v_h\|_{L^p(D)} = 0$ for all $v \in L^p(D)$. (*Hint*: by density.)

Exercise 19.5 (Hermite). Let $\mathcal{T}_h := \{[x_i, x_{i+1}]\}_{i\in\{0:I\}}$ be a mesh of the interval $D := (a, b)$. Recall the Hermite finite element from Exercise 5.4. Specify global shape functions $\{\varphi_{i,0}, \varphi_{i,1}\}_{i\in\{0:I+1\}}$ in $H_h := \{v_h \in C^1(\overline{D}) \mid \forall i \in \{0:I\},\, v_{h|[x_i,x_{i+1}]} \in \mathbb{P}_3\}$. (*Hint*: consider values of the function or of its derivative at the mesh nodes.) Can the bicubic Hermite rectangular finite element from Exercise 6.8 be used to enforce C^1-continuity for $d = 2$?

Solution to exercises

Exercise 19.1 (Connectivity classes). (i) We have

$$\texttt{j\_dof} : \{1{:}5\}\times\{1{:}6\} \to \{1{:}18\}.$$

(ii) Recall that $(K, i) \in \texttt{j\_dof}^{-1}(a)$ iff $\texttt{j\_dof}(K, i) = a$. Hence, we have

$$\begin{aligned}\texttt{j\_dof}^{-1}(8) &= \{(3, 3), (2, 3), (4, 3)\},\\ \texttt{j\_dof}^{-1}(13) &= \{(5, 4), (3, 6)\}.\end{aligned}$$

(iii) Recall that $\mathcal{T}_a := \{K \in \mathcal{T}_h \mid \exists i \in \mathcal{N},\, \texttt{j\_dof}(K, i) = a\}$. Hence, we have

$$\begin{aligned}\mathcal{T}_6 &= \{K_1, K_2, K_3, K_5\},\\ \mathcal{T}_{10} &= \{K_2, K_4\}.\end{aligned}$$

Exercise 19.2 (Stiffness, mass, incidence matrices). (i) Let us first notice that $\nabla\lambda_n \in \boldsymbol{P}_0^{\mathrm{c}}(\mathcal{T}_h)$ for all $n \in \{1{:}N_{\mathrm{v}}\}$. Since $\{\sigma_m^{\mathrm{e}}\}_{m\in\{1:N_{\mathrm{e}}\}}$ is the dual basis of $\{\boldsymbol{\theta}_m\}_{m\in\{1:N_{\mathrm{e}}\}}$, the assertion is proved by showing that $\sigma_m^{\mathrm{e}}(\nabla\lambda_n) = \mathcal{M}_{mn}^{\mathrm{ev}}$ for all $n \in \{1{:}N_{\mathrm{v}}\}$ and all $m \in \{1{:}N_{\mathrm{e}}\}$. We have

$$\sigma_m^{\mathrm{e}}(\nabla\lambda_n) = \frac{1}{|E_m|}\int_{E_m} (\nabla\lambda_n)\cdot\boldsymbol{t}_{E_m}\,\mathrm{d}l,$$

where $\boldsymbol{t}_{E_m}$ is the vector orienting E_m (recall that $\|\boldsymbol{t}_{E_m}\|_{\ell^2} = |E_m|$). Let $\{\boldsymbol{z}_p, \boldsymbol{z}_q\}$ be the two endpoints of E_m so that $\boldsymbol{t}_{E_m}$ points from $\boldsymbol{z}_p$ to $\boldsymbol{z}_q$. We have

$$\sigma_m^{\mathrm{e}}(\nabla\lambda_n) = \lambda_n(\boldsymbol{z}_q) - \lambda_n(\boldsymbol{z}_p) = \delta_{nq} - \delta_{np} = \mathcal{M}_{mn}^{\mathrm{ev}},$$

by definition of the incidence matrix $\mathcal{M}^{\mathrm{ev}}$. This completes the proof.
(ii) Using that $\nabla\lambda_n = \sum_{m\in\{1:N_{\mathrm{e}}\}} \mathcal{M}_{mn}^{\mathrm{ev}}\boldsymbol{\theta}_m$ for all $n \in \{1{:}N_{\mathrm{v}}\}$, we infer that for all $n, n' \in \{1{:}N_{\mathrm{v}}\}$,

$$\begin{aligned}\mathcal{A}_{nn'} &= \int_D \nabla\lambda_n\cdot\nabla\lambda_{n'}\,\mathrm{d}x = \sum_{m\in\{1:N_{\mathrm{e}}\}}\sum_{m'\in\{1:N_{\mathrm{e}}\}} \mathcal{M}_{mn}^{\mathrm{ev}}\Big(\int_D \boldsymbol{\theta}_m\cdot\boldsymbol{\theta}_{m'}\,\mathrm{d}x\Big)\mathcal{M}_{m'n'}^{\mathrm{ev}}\\ &= \sum_{m\in\{1:N_{\mathrm{e}}\}}\sum_{m'\in\{1:N_{\mathrm{e}}\}} \mathcal{M}_{mn}^{\mathrm{ev}}\mathcal{N}_{mm'}\mathcal{M}_{m'n'}^{\mathrm{ev}} = \big((\mathcal{M}^{\mathrm{ev}})^{\mathsf{T}}\mathcal{N}\mathcal{M}^{\mathrm{ev}}\big)_{nn'}.\end{aligned}$$

This proves the expected identity.

Exercise 19.3 (Zero trace). (i) Let $a \in \mathcal{A}_h^{\circ}$. For all $a' \in \mathcal{A}_h^{\partial}$, we have $\sigma_{a'}(\varphi_a) = \delta_{aa'} = 0$ because $\{\mathcal{A}_h^{\partial}, \mathcal{A}_h^{\circ}\}$ forms a partition of $\mathcal{A}_h$. We conclude by invoking (19.38a) and the definition of $P_{k,0}^{\mathrm{x}}(\mathcal{T}_h)$.

(ii) We have already established that the set $\{\varphi_a\}_{a\in\mathcal{A}_h}$ is linearly independent. Hence, $\{\varphi_a\}_{a\in\mathcal{A}_h^\partial}$ is also linearly independent. For all $v \in P_{k,0}^{\mathrm{x}}(\mathcal{T}_h) \subset P_k^{\mathrm{x}}(\mathcal{T}_h)$, we have

$$v = \sum_{a\in\mathcal{A}_h^\partial} \sigma_a(v)\varphi_a + \sum_{a\in\mathcal{A}_h^\circ} \sigma_a(v)\varphi_a,$$

but by definition $\sigma_a(v) = 0$ for all $a \in \mathcal{A}_h^\partial$ (see Definition 19.11). Hence, $v = \sum_{a\in\mathcal{A}_h^\circ} \sigma_a(v)\varphi_a$, thereby showing that $\{\varphi_a\}_{a\in\mathcal{A}_h^\circ}$ is a spanning set.

Exercise 19.4 (Approximability in L^p). Let $\epsilon > 0$. Let l be as in Corollary 19.8 and set $s := \max(l, k+1)$. Since $W^{s,p}(D)$ is dense in $L^p(D)$, there is $v_\epsilon \in W^{s,p}(D)$ such that $\|v - v_\epsilon\|_{L^p(D)} \le \epsilon$. Since $\mathcal{I}_h^{\mathrm{L}}(v_\epsilon) \in P_k^{\mathrm{g}}(\mathcal{T}_h)$, the triangle inequality gives

$$\begin{aligned}\inf_{v_h\in P_k^{\mathrm{g}}(\mathcal{T}_h)} \|v - v_h\|_{L^p(D)} &\le \|v - \mathcal{I}_h^{\mathrm{L}}(v_\epsilon)\|_{L^p(D)} \\ &\le \|v - v_\epsilon\|_{L^p(D)} + \|v_\epsilon - \mathcal{I}_h^{\mathrm{L}}(v_\epsilon)\|_{L^p(D)}.\end{aligned}$$

Owing to Corollary 19.8 with $m := 0$, we infer that the second term tends to zero as $h \to 0$. Hence, $\limsup_{h\to 0}(\inf_{v_h\in P_k^{\mathrm{g}}(\mathcal{T}_h)} \|v - v_h\|_{L^p(D)}) \le \epsilon$, and the conclusion follows since ϵ is arbitrary.

Exercise 19.5 (Hermite). Let $\{\widehat{\theta}_i\}_{i\in\{1:4\}}$ be the shape functions of the Hermite finite element on the reference interval $[0,1]$; see Exercise 5.4. This yields

$$\varphi_{i,0}(x) = \begin{cases} \widehat{\theta}_3\left(\frac{x-x_{i-1}}{x_i-x_{i-1}}\right) & \text{if } x \in [x_{i-1}, x_i], \\ \widehat{\theta}_1\left(\frac{x-x_i}{x_{i+1}-x_i}\right) & \text{if } x \in [x_i, x_{i+1}], \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\varphi_{i,1}(x) = \begin{cases} h_{i-1}\widehat{\theta}_4\left(\frac{x-x_{i-1}}{x_i-x_{i-1}}\right) & \text{if } x \in [x_{i-1}, x_i], \\ h_i\widehat{\theta}_2\left(\frac{x-x_i}{x_{i+1}-x_i}\right) & \text{if } x \in [x_i, x_{i+1}], \\ 0 & \text{otherwise,} \end{cases}$$

with $h_i := x_{i+1} - x_i$ for all $i \in \{0{:}I\}$. Proceeding as in the proof of Proposition 19.4 shows that these global shape functions are linearly independent and form a spanning set of the whole space H_h. Finally, the bicubic Hermite rectangular finite element can indeed be used to enforce C^1-continuity. For instance, consider $p \in \mathbb{Q}_{3,2}$. On the face $\{x_1 = 1\}$, the x_2-dependent function $\partial_{x_1} p|_{\{x_1=1\}}$ is in $\mathbb{P}_{3,1}$. Owing to the choice of the dofs, its values and x_2-derivatives are the same on both sides of the face.

Chapter 20

Face gluing

Exercises

Exercise 20.1 (Affine mapping between faces). Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$ and set $\widehat{F}_l := \boldsymbol{T}_{K_l}^{-1}(F)$ and $\widehat{F}_r := \boldsymbol{T}_{K_r}^{-1}(F)$. Prove that the mapping $\boldsymbol{T}_{rl} := \boldsymbol{T}_{K_l}^{-1} \circ \boldsymbol{T}_{K_r|\widehat{F}_r}$ is affine. (*Hint*: let $(\widehat{K}, \widehat{P}_{\mathrm{geo}}, \widehat{\Sigma}_{\mathrm{geo}})$ be the geometric reference Lagrange finite element. Observe that the two face finite elements $(\widehat{F}_l, \widehat{P}^{\mathrm{g}}_{\mathrm{geo},l}, \widehat{\Sigma}^{\mathrm{g}}_{\mathrm{geo},l})$ and $(\widehat{F}_r, \widehat{P}^{\mathrm{g}}_{\mathrm{geo},r}, \widehat{\Sigma}^{\mathrm{g}}_{\mathrm{geo},r})$ can be constructed from the same reference Lagrange finite element $(\widehat{F}^{d-1}, \widehat{P}^{d-1}_{\mathrm{geo}}, \widehat{\Sigma}^{d-1}_{\mathrm{geo}})$.)

Exercise 20.2 (Linear maps). Let E, F, G be finite-dimensional vector spaces, let $A \in \mathcal{L}(E;F)$ and let $T \in \mathcal{L}(E;G)$. Assume that $\ker(T) \subset \ker(A)$. Set $\tilde{G} := T(E)$. (i) Prove that there is $\tilde{A} \in \mathcal{L}(\tilde{G};F)$ s.t. $A = \tilde{A} \circ T$. (*Hint*: build a right inverse of T using a direct sum $E = E_1 \oplus E_2$ with $E_1 := \ker(T)$.) (ii) Show that $\tilde{A}$ is uniquely defined, i.e., does not depend on E_2.

Exercise 20.3 ($\gamma_{K,F}$ and $\mathcal{N}_{K,F}$). (i) Prove that $P_K = \sum_{F\in\mathcal{F}_K} \ker(\gamma^{\mathrm{x}}_{K,F})$ (nondirect sum of vector spaces) if and only if there is $F \in \mathcal{F}_K$ s.t. $i \notin \mathcal{N}_{K,F}$ for all $i \in \mathcal{N}$. (ii) Let the face unisolvence assumption hold true. Let $\mathcal{F}(K,i) := \{F \in \mathcal{F}_K \mid \ker(\gamma_{K,F}) \subset \ker(\sigma_{K,i})\}$. Prove the following statements: (ii.a) $F \in \mathcal{F}(K,i)$ iff $i \in \mathcal{N}_{K,F}$; (ii.b) $F \in \mathcal{F}(K,i)$ iff $\gamma_{K,F}(\theta_{K,i}) \neq 0$ where $\theta_{K,i}$ is the local shape function associated with the dof i.

Exercise 20.4 (Reference face element). Let $\widehat{F}$ be any face of $\widehat{K}$. Let $\widehat{P}^{\mathrm{x}} := \gamma^{\mathrm{x}}_{\widehat{K},\widehat{F}}(\widehat{P})$ and let $\mathcal{N}_{\widehat{K},\widehat{F}}$ be the subset of $\mathcal{N}$ s.t. $\bigcap_{i\in\mathcal{N}_{\widehat{K},\widehat{F}}} \ker(\sigma_{\widehat{K},i}) = \ker(\gamma_{\widehat{K},\widehat{F}})$. Recall that this means that there exists $\widehat{\sigma}^{\mathrm{x}}_{\widehat{F},i} : \widehat{P}_{\widehat{K},\widehat{F}} \to \mathbb{R}$ s.t. $\widehat{\sigma}_i = \widehat{\sigma}^{\mathrm{x}}_{\widehat{F},i} \circ \gamma^{\mathrm{x}}_{\widehat{K},\widehat{F}}$ for all $i \in \mathcal{N}_{\widehat{K},\widehat{F}}$. Assume that $\mathcal{N}_{\widehat{K},\widehat{F}}$ is nonempty, that the triple $\{\widehat{F}, \widehat{P}^{\mathrm{x}}, \widehat{\Sigma}^{\mathrm{x}}\}$ with $\widehat{\Sigma}^{\mathrm{x}} := \{\widehat{\sigma}^{\mathrm{x}}_{\widehat{F},i}\}_{i\in\mathcal{N}_{\widehat{K},\widehat{F}}}$ is a finite element, and that there is a linear bijective map $\psi_F : P^{\mathrm{x}}_{K,F} \to \widehat{P}^{\mathrm{x}}$ s.t. $\psi_F^{-1} \circ \gamma^{\mathrm{x}}_{\widehat{K},\widehat{F}} = \gamma^{\mathrm{x}}_{K,F} \circ \psi_K^{-1}$. Prove that Assumption 20.12 holds true and $\mathcal{N}_{K,F} = \mathcal{N}_{\widehat{K},\widehat{F}}$. (*Hint*: show that the finite element $\{F, P^{\mathrm{x}}_{K,F}, \Sigma^{\mathrm{x}}_{K,F}\}$ is generated from $\{\widehat{F}, \widehat{P}^{\mathrm{x}}, \widehat{\Sigma}^{\mathrm{x}}\}$ using the map ψ_F.)

Exercise 20.5 (Permutation invariance). Let $\widehat{S}^1 := [0,1]$ and consider the bases $\mathfrak{B}_1 := \{\mu_1(s) = 1-s, \mu_2(s) = s\}$ and $\mathfrak{B}_2 := \{\mu_1(s) = 1, \mu_2(s) = s\}$. Are these bases invariant under permutation of the vertices of $\widehat{S}^1$?

Exercise 20.6 (Canonical hybrid element, $d=3$). Consider the assumptions made in §20.4.3. (i) Prove the face unisolvence assumption 20.12. (ii) Let $F \in \mathcal{F}_K$. Let $\boldsymbol{T}_{\widehat{F}} : \widehat{S}^2 \to \widehat{F}$ be an affine bijective mapping, and let $\boldsymbol{T}_{K,F} := \boldsymbol{T}_{K|\widehat{F}} \circ \boldsymbol{T}_{\widehat{F}} : \widehat{S}^2 \to F$. Verify that $P^{\mathrm{g}}_{K,F} = \mathbb{P}_{k,d-1} \circ \boldsymbol{T}_{K,F}^{-1}$ and that $\{F, P^{\mathrm{g}}_{K,F}, \Sigma^{\mathrm{g}}_{K,F}\}$ is a two-dimensional canonical hybrid element. (iii) Prove that $P^{\mathrm{g}}_{K_l,F} = P^{\mathrm{g}}_{K_r,F} =: P^{\mathrm{g}}_F$ and $\Sigma^{\mathrm{g}}_{K_l,F} = \Sigma^{\mathrm{g}}_{K_r,F} =: \Sigma^{\mathrm{g}}_F$.

Exercise 20.7 ($P_{K,F}$). Let $\widehat{K}$ be the unit simplex in $\mathbb{R}^2$ and let $\{\widehat{F}_i\}_{i\in\{0:2\}}$ be the faces of $\widehat{K}$. Recall that for $\mathbb{P}_{k,d}$ scalar-valued elements, we have $P_{\widehat{K},\widehat{F}_i} := \gamma^{\mathrm{g}}_{\widehat{K},\widehat{F}_i}(\mathbb{P}_{k,d})$. (i) Compute a basis of $P_{\widehat{K},\widehat{F}_i}$ for all $i \in \{0:2\}$ assuming that $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ is the $\mathbb{P}_1$ Lagrange element. Is $(\widehat{F}_i, P_{\widehat{K},\widehat{F}_i}, \Sigma_{\widehat{K},\widehat{F}_i})$ a finite element? (ii) Compute a basis of $P_{\widehat{K},\widehat{F}_i}$ for all $i \in \{0:2\}$ assuming that $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ is the $\mathbb{P}_1$ Crouzeix–Raviart element. Is $(\widehat{F}_i, P_{\widehat{K},\widehat{F}_i}, \Sigma_{\widehat{K},\widehat{F}_i})$ a finite element?

Solution to exercises

Exercise 20.1 (Affine mapping between faces). Let $(\widehat{K}, \widehat{P}_{\mathrm{geo}}, \widehat{\Sigma}_{\mathrm{geo}})$ be the geometric reference Lagrange finite element. By assumption, the two face finite elements

$$(\widehat{F}_l, \widehat{P}^{\mathrm{g}}_{\mathrm{geo},l}, \widehat{\Sigma}^{\mathrm{g}}_{\mathrm{geo},l}) \quad \text{and} \quad (\widehat{F}_r, \widehat{P}^{\mathrm{g}}_{\mathrm{geo},r}, \widehat{\Sigma}^{\mathrm{g}}_{\mathrm{geo},r})$$

can be constructed from the same reference Lagrange finite element $(\widehat{F}^{d-1}, \widehat{P}^{d-1}_{\mathrm{geo}}, \widehat{\Sigma}^{d-1}_{\mathrm{geo}})$. Let $\{\widehat{\theta}_n\}_{n\in\mathcal{N}^{d-1}}$ be the reference shape functions of $(\widehat{F}^{d-1}, \widehat{P}^{d-1}_{\mathrm{geo}}, \widehat{\Sigma}^{d-1}_{\mathrm{geo}})$, and let $\{\widehat{\psi}_n\}_{n\in\mathcal{N}_{\mathrm{geo}}}$ be the reference shape functions of $(\widehat{K}, \widehat{P}_{\mathrm{geo}}, \widehat{\Sigma}_{\mathrm{geo}})$. Let $\mathcal{N}_{\mathrm{geo},l}$ and $\mathcal{N}_{\mathrm{geo},r}$ be the indices of the geometric Lagrange nodes from K_l and K_r on F. These two sets of nodes must be identical, i.e., there exist two bijective maps $j_l : \mathcal{N}^{d-1} \to \mathcal{N}_{\mathrm{geo},l}$ and $j_r : \mathcal{N}^{d-1} \to \mathcal{N}_{\mathrm{geo},r}$ such that $\boldsymbol{g}_{\mathtt{j\_geo}(j_l(n),K_l)} = \boldsymbol{g}_{\mathtt{j\_geo}(j_r(n),K_r)}$ and $\widehat{\theta}_n = \widehat{\psi}_{j_l(n)} \circ \boldsymbol{T}_{\widehat{F}_l} = \widehat{\psi}_{j_r(n)} \circ \boldsymbol{T}_{\widehat{F}_r}$ for all $n \in \mathcal{N}^{d-1}$, where $\boldsymbol{T}_{\widehat{F}_l} : \widehat{F}^{d-1} \to \widehat{F}_l$ and $\boldsymbol{T}_{\widehat{F}_r} : \widehat{F}^{d-1} \to \widehat{F}_r$ are the two affine geometric mappings which map the vertices of $\widehat{F}^{d-1}$ to the vertices of $\widehat{F}_l$ and $\widehat{F}_r$, respectively. We have for all $\widehat{\boldsymbol{x}}$ in $\widehat{F}^{d-1}$,

$$\begin{aligned}
\boldsymbol{T}_{K_l|F_l}(\boldsymbol{T}_{\widehat{F}_l}(\widehat{\boldsymbol{x}})) &= \sum_{m\in\mathcal{N}_{\mathrm{geo},l}} \boldsymbol{g}_{\mathtt{j_geo}(m,K_l)} \widehat{\psi}_m(\boldsymbol{T}_{\widehat{F}_l}(\widehat{\boldsymbol{x}})) \\
&= \sum_{n\in\mathcal{N}^{d-1}} \boldsymbol{g}_{\mathtt{j_geo}(j_l(n),K_l)} \widehat{\psi}_{j_l(n)}(\boldsymbol{T}_{\widehat{F}_l}(\widehat{\boldsymbol{x}})) \\
&= \sum_{n\in\mathcal{N}^{d-1}} \boldsymbol{g}_{\mathtt{j_geo}(j_r(n),K_r)} \widehat{\psi}_{j_r(n)}(\boldsymbol{T}_{\widehat{F}_r}(\widehat{\boldsymbol{x}})) = \boldsymbol{T}_{K_r|F_r}(\boldsymbol{T}_{\widehat{F}_r}(\widehat{\boldsymbol{x}})).
\end{aligned}$$

This proves that $\boldsymbol{T}_{K_l|F_l} \circ \boldsymbol{T}_{\widehat{F}_l} = \boldsymbol{T}_{K_r|F_r} \circ \boldsymbol{T}_{\widehat{F}_r}$, i.e., $\boldsymbol{T}^{-1}_{K_l|F_l} \circ \boldsymbol{T}_{K_r|F_r} = \boldsymbol{T}_{\widehat{F}_l} \circ \boldsymbol{T}^{-1}_{\widehat{F}_r}$. Hence, $\boldsymbol{T}^{-1}_{K_l|F_l} \circ \boldsymbol{T}_{K_r|F_r}$ is affine since $\boldsymbol{T}_{\widehat{F}_l} \circ \boldsymbol{T}^{-1}_{\widehat{F}_r}$ is affine.

Exercise 20.2 (Linear maps). (i) Let $E = E_1 \oplus E_2$ be one direct-sum decomposition of E with $E_1 := \ker(T)$ (this is always possible since E is finite-dimensional). For all $x \in E$, we write $x = x_1 + x_2$ with $x_1 \in E_1 := \ker(T)$ and $x_2 \in E_2$. Let $\tilde{T} : E_2 \to T(E)$ be such that $\tilde{T}(e_2) = T(e_2)$ for all $e_2 \in E_2$. Let $e_2 \in E_2$ be such that $\tilde{T}(e_2) = 0$. Then $e_2 \in \ker(T) \cap E_2 = E_1 \cap E_2 = \{0\}$, whence $e_2 = 0$. This proves that $\tilde{T}$ is injective. Let $\tilde{g} \in T(E)$. There is $e = e_1 + e_2 \in E$ such that $T(e) = \tilde{g}$. Hence, $\tilde{g} = T(e) = T(e_2) = \tilde{T}(e_2)$. This proves that $\tilde{T}$ is surjective. In conclusion, $\tilde{T}$ is

bijective. Note that $T\circ\tilde{T}^{-1}(\tilde{g})=\tilde{T}\circ\tilde{T}^{-1}(\tilde{g})=\tilde{g}$ for all $\tilde{g}\in T(E)$ since $\tilde{T}^{-1}(\tilde{g})\in E_2$. Hence, $\tilde{T}^{-1}$ is a right inverse of T. Set $\tilde{A}:=A\circ\tilde{T}^{-1}:T(E)\to F$. Using that $A(x_1)=0$ for all $x_1\in E_1$ since $\ker(T)\subset\ker(A)$, we infer that

$$\begin{aligned}(\tilde{A}\circ T)(x)&=(A\circ\tilde{T}^{-1}\circ T)(x)=(A\circ\tilde{T}^{-1})(T(x_2))=(A\circ\tilde{T}^{-1}\circ\tilde{T})(x_2)\\&=A(x_2)=A(x_1+x_2)=A(x).\end{aligned}$$

Hence, $\tilde{A}\circ T=A$.
(ii) Let us show that $\tilde{A}$ is uniquely defined, i.e., $\tilde{A}$ does not depend on the choice of E_2 in the direct sum $E=E_1\oplus E_2$. Let $\tilde{A}_1:T(E)\to F$ and $\tilde{A}_2:T(E)\to F$ be two maps constructed as above using two different subspaces E_2. We have $\tilde{A}_1\circ T=A=\tilde{A}_2\circ T$. Let $\tilde{g}\in T(E)$. Thus, there is $e\in E$ such that $T(e)=\tilde{g}$. This implies that

$$\tilde{A}_1(\tilde{g})=\tilde{A}_1(T(e))=(\tilde{A}_1\circ T)(e)=(\tilde{A}_2\circ T)(e)=\tilde{A}_2(T(e))=\tilde{A}_2(\tilde{g}).$$

Hence, $\tilde{A}_1=\tilde{A}_2$.

Exercise 20.3 ($\gamma_{K,F}$ and $\mathcal{N}_{K,F}$). (i) Assume that $P_K=\sum_{F\in\mathcal{F}_K}\ker(\gamma^{\mathrm{x}}_{K,F})$. Let us reason by contradiction and assume that there is $i\in\mathcal{N}$ such that $i\in\mathcal{N}_{K,F}$ for all $F\in\mathcal{F}_K$. Since $i\in\mathcal{N}_{K,F}$ implies that $\ker(\gamma^{\mathrm{x}}_{K,F})\subset\ker(\sigma_{K,i})$, and since this inclusion holds true for all $F\in\mathcal{F}_K$, we obtain $P_K=\sum_{F\in\mathcal{F}_K}\ker(\gamma^{\mathrm{x}}_{K,F})\subset\ker(\sigma_{K,i})$, which contradicts that $\sigma_{K,i}(\theta_{K,i})=1$. Conversely, assume that for all $i\in\mathcal{N}$, there is $F\in\mathcal{F}_K$ s.t. $i\notin\mathcal{N}_{K,F}$, that is, $\theta_{K,i}\in\ker(\gamma^{\mathrm{x}}_{K,F})$. Since any function in P_K can be written as a linear combination of the functions $\theta_{K,i}$, we conclude that $P_K=\sum_{F\in\mathcal{F}_K}\ker(\gamma^{\mathrm{x}}_{K,F})$.
(ii) Let us assume that the face unisolvence assumption holds true.
(ii.a) Let $F\in\mathcal{F}(K,i)$ and assume that i is not in $\mathcal{N}_{K,F}$. This implies that $\sigma_{K,j}(\theta_{K,i})=\delta_{ij}=0$ for all $j\in\mathcal{N}_{K,F}$ because $i\notin\mathcal{N}_{K,F}$. Recall that the face unisolvence assumption (Assumption (20.12)) says that $\ker(\gamma_{K,F})=\bigcap_{j\in\mathcal{N}_{K,F}}\ker(\sigma_{K,j})$. Hence, we have $\theta_{K,i}\in\ker(\gamma_{K,F})$, which, in turn, implies that $\theta_{K,i}\in\ker(\sigma_{K,i})$ because $F\in\mathcal{F}(K,i)$. This is absurd. Hence, $i\in\mathcal{N}_{K,F}$. Let us assume now that $i\in\mathcal{N}_{K,F}$. Then $\ker(\gamma_{K,F})=\bigcap_{j\in\mathcal{N}_{K,F}}\ker(\sigma_{K,j})\subset\ker(\sigma_{K,i})$, which implies that $F\in\mathcal{F}(K,i)$.
(ii.b) Let $F\in\mathcal{F}(K,i)$ and assume that $\gamma_{K,F}(\theta_{K,i})=0$. Then $\sigma_{K,i}(\theta_{K,i})=0$ because $F\in\mathcal{F}(K,i)$, which is absurd. Hence, $\gamma_{K,F}(\theta_{K,i})\neq0$. Assume now that $\gamma_{K,F}(\theta_{K,i})\neq0$, and assume that $F\notin\mathcal{F}(K,i)$, which owing to the above characterization of $\mathcal{F}(k,i)$ means that $i\notin\mathcal{N}_{K,i}$. Then $\sigma_{K,j}(\theta_{K,i})=\delta_{ij}=0$ for all $j\in\mathcal{N}_{K,F}$, which owing to the face unisolvence assumption implies that $\gamma_{K,F}(\theta_{K,i})=0$ which is absurd. Hence, $F\in\mathcal{F}(K,i)$.

Exercise 20.4 (Reference face element). The fact that $\widehat{P}^{\mathrm{x}}=\psi_F(P^{\mathrm{x}}_{K,F})$ follows from

$$\widehat{P}^{\mathrm{x}}=\gamma^{\mathrm{x}}_{\widehat{K},\widehat{F}}(\widehat{P})=\psi_F(\gamma^{\mathrm{x}}_{K,F}(\psi_K^{-1}(\widehat{P})))=\psi_F(\gamma^{\mathrm{x}}_{K,F}(P_K))=\psi_F(P^{\mathrm{x}}_{K,F}).$$

Let us now show that $\mathcal{N}_{\widehat{K},\widehat{F}}=\mathcal{N}_{K,F}$. Indeed, $i\in\mathcal{N}_{\widehat{K},\widehat{F}}$ means that there is $\widehat{\sigma}^{\mathrm{x}}_{\widehat{F},i}:\widehat{P}^{\mathrm{x}}\to\mathbb{R}$ s.t. $\widehat{\sigma}_i=\widehat{\sigma}^{\mathrm{x}}_{\widehat{F},i}\circ\gamma^{\mathrm{x}}_{\widehat{K},\widehat{F}}$. Defining $\sigma^{\mathrm{x}}_{K,F,i}:P^{\mathrm{x}}_{K,F}\to\mathbb{R}$ by $\sigma^{\mathrm{x}}_{K,F,i}:=\widehat{\sigma}^{\mathrm{x}}_{\widehat{F},i}\circ\psi_F$, we obtain for all $p\in P_K$,

$$\begin{aligned}\sigma_{K,i}(p)&=\widehat{\sigma}_i(\psi_K(p))=\widehat{\sigma}^{\mathrm{x}}_{\widehat{F},i}(\gamma^{\mathrm{x}}_{\widehat{K},\widehat{F}}(\psi_K(p)))\\&=\widehat{\sigma}^{\mathrm{x}}_{\widehat{F},i}(\psi_F(\gamma^{\mathrm{x}}_{K,F}(p)))=\sigma^{\mathrm{x}}_{K,F,i}(\gamma^{\mathrm{x}}_{K,F}(p)),\end{aligned}$$

so that $i\in\mathcal{N}_{K,F}$. This proves that $\mathcal{N}_{\widehat{K},\widehat{F}}\subset\mathcal{N}_{K,F}$, and the converse inclusion is proved similarly. Since $\widehat{\sigma}_i=\widehat{\sigma}^{\mathrm{x}}_{\widehat{F},i}\circ\gamma^{\mathrm{x}}_{\widehat{K},\widehat{F}}$, we conclude that $\{F,P^{\mathrm{x}}_{K,F},\Sigma^{\mathrm{x}}_{K,F}\}$ is generated from $\{\widehat{F},\widehat{P}^{\mathrm{x}},\widehat{\Sigma}^{\mathrm{x}}\}$ using the map ψ_F.

Exercise 20.5 (Permutation invariance). There are two possible permutations of the vertices, the mappings $S_1(s) = s$ (under which the vertices are invariant, and thus invariance of any basis is trivial) and $S_2(s) = 1 - s$ (exchanging the two vertices). The basis $\mathfrak{B}_1$ is left invariant by the permutation S_2 since $\mu_1 \circ S_2 = \mu_2$ and $\mu_2 \circ S_2 = \mu_1$. This is not the case for the basis $\mathfrak{B}_2$ for which $\mu_2 \circ S_2$ differs from μ_1 and from μ_2.

Exercise 20.6 (Canonical hybrid element, $d = 3$). (i) Let $F \in \mathcal{F}_K$. Let

$$\mathcal{N}_{K,F}^{\mathrm{v}} := \{i \in \mathcal{N} \mid \exists \boldsymbol{z}(i) \in \mathcal{V}_F,\ \sigma_{K,i} = \sigma_{\boldsymbol{z}(i)}^{\mathrm{v}}\},$$

be the collection of the vertex dofs associated with F, let (if $k \geq 2$)

$$\mathcal{N}_{K,F}^{\mathrm{e}} := \{i \in \mathcal{N} \mid \exists (E(i), m(i)) \in \mathcal{E}_F \times \{1{:}n_{\mathrm{sh}}^{\mathrm{e}}\},\ \sigma_{K,i} = \sigma_{E(i),m(i)}^{\mathrm{e}}\},$$

be the collection of the edge dofs associated with F, and let (if $k \geq 3$)

$$\mathcal{N}_{K,F}^{\mathrm{f}} := \{i \in \mathcal{N} \mid \exists m(i) \in \{1{:}n_{\mathrm{sh}}^{\mathrm{f}}\},\ \sigma_{K,i} = \sigma_{F,m(i)}^{\mathrm{f}}\}.$$

We adopt the convention $\mathcal{N}_{K,F}^{\mathrm{e}} := \emptyset$ if $k = 1$ and $\mathcal{N}_{K,F}^{\mathrm{f}} := \emptyset$ if $k \leq 2$. Let us set $\mathcal{N}_{K,F} := \mathcal{N}_{K,F}^{\mathrm{v}} \cup \mathcal{N}_{K,F}^{\mathrm{e}} \cup \mathcal{N}_{K,F}^{\mathrm{f}}$. We first observe that the set $\mathcal{N}_{K,F}$ is nonempty. Moreover, since $\gamma_{K,F}^{\mathrm{g}}(v) := v_{|F}$, we infer that $\gamma_{K,F}^{\mathrm{g}}(v) = 0$ implies that $\sigma_{K,i}(v) = 0$ for all $i \in \mathcal{N}_{K,F}$, i.e., $\ker(\gamma_{K,F}^{\mathrm{g}}) \subset \bigcap_{i \in \mathcal{N}_{K,F}} \ker(\sigma_{K,i})$. The converse inclusion follows from the proof of Proposition 7.19.
(ii) Let $F \in \mathcal{F}_K$. We have already shown in Lemma 20.5 that $P_{K,F}^{\mathrm{g}} := \gamma_{K,F}^{\mathrm{g}}(P_K) = \mathbb{P}_{k,d-1} \circ \boldsymbol{T}_{K,F}^{-1}$. Moreover, let us consider the following linear forms:

$$\begin{aligned}
\sigma_{K,F,i}^{\mathrm{v}}(v) &:= v(\boldsymbol{z}(i)), && \forall i \in \mathcal{N}_{K,F}^{\mathrm{v}},\\
\sigma_{K,F,i}^{\mathrm{e}}(v) &:= \frac{1}{|E(i)|} \int_{E(i)} (\mu_{m(i)} \circ \boldsymbol{T}_{K,E(i)}^{-1}) v \,\mathrm{d}l, && \forall i \in \mathcal{N}_{K,F}^{\mathrm{e}},\\
\sigma_{K,F,i}^{\mathrm{f}}(v) &:= \frac{1}{|F|} \int_{F} (\zeta_{m(i)} \circ \boldsymbol{T}_{K,F}^{-1}) v \,\mathrm{d}s, && \forall i \in \mathcal{N}_{K,F}^{\mathrm{s}},
\end{aligned}$$

where $\boldsymbol{T}_{K,E(i)} := \boldsymbol{T}_{K|\widehat{E}(i)} \circ \boldsymbol{T}_{\widehat{E}(i)} : \widehat{S}^1 \to E(i)$, $\widehat{E}(i) := \boldsymbol{T}_K^{-1}(E(i))$, and $\boldsymbol{T}_{\widehat{E}(i)} : \widehat{S}^1 \to \widehat{E}(i)$, and similarly $\boldsymbol{T}_{K,F} = \boldsymbol{T}_{K|\widehat{F}} \circ \boldsymbol{T}_{\widehat{F}} : \widehat{S}^2 \to F$, $\widehat{F} := \boldsymbol{T}_K^{-1}(F)$, and $\boldsymbol{T}_{\widehat{F}} : \widehat{S}^2 \to \widehat{F}$. Since for all $i \in \mathcal{N}_{K,F}$ and all x $\in$ {v,e,f}, we have $\sigma_{K,i}^{\mathrm{x}}(v) = \sigma_{K,F,i}^{\mathrm{x}}(\gamma_{K,F}^{\mathrm{g}}(v))$, the definition (20.10) implies that the set $\Sigma_{K,F}^{\mathrm{g}}$ is exactly the above collection of dofs. Moreover, these expressions show that the triple $\{F, P_{K,F}^{\mathrm{g}}, \Sigma_{K,F}^{\mathrm{g}}\}$ is a two-dimensional canonical hybrid element.
(iii) Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^{\circ}$. We have already shown in Lemma 20.6 that $P_{K_l,F}^{\mathrm{g}} = P_{K_r,F}^{\mathrm{g}} =: P_F^{\mathrm{g}}$. To prove that $\Sigma_{K_l,F}^{\mathrm{g}} = \Sigma_{K_r,F}^{\mathrm{g}} =: \Sigma_F^{\mathrm{g}}$, we have to construct a bijective map $\chi_{lr} : \mathcal{N}_{K_l,F} \to \mathcal{N}_{K_r,F}$ such that $\sigma_{K_r,F,\chi_{lr}(i)}^{\mathrm{g}} = \sigma_{K_l,F,i}^{\mathrm{g}}$ for all $i \in \mathcal{N}_{K_l,F}$. Let $i \in \mathcal{N}_{K_l,F}$. We distinguish three cases.
(iii.a) Assume that $i \in \mathcal{N}_{K_l,F}^{\mathrm{v}}$. Then $\boldsymbol{z}(i)$ is a vertex of F so that there is $i_r \in \mathcal{N}_{K_r,F}^{\mathrm{v}}$ such that $\boldsymbol{z}(i) = \boldsymbol{z}(i_r)$. We set $\chi_{lr}(i) := i_r$, and this gives $\sigma_{K_r,F,\chi_{lr}(i)}^{\mathrm{g}}(v) = \sigma_{K_l,F,i}^{\mathrm{g}}(v)$ for all $v \in P_F^{\mathrm{g}}$. Notice that $\chi_{lr}(\mathcal{N}_{K_l,F}^{\mathrm{v}}) = \mathcal{N}_{K_r,F}^{\mathrm{v}}$.
(iii.b) Assume $i \in \mathcal{N}_{K_l,F}^{\mathrm{e}}$ and consider the associated pair $(E(i), m(i)) \in \mathcal{E}_F \times \{1{:}n_{\mathrm{sh}}^{\mathrm{e}}\}$. Since $E(i)$ is an edge of F which is a face of K_l an K_r, we can consider the mappings $\boldsymbol{T}_{K_l,E(i)} : \widehat{S}^1 \to E(i)$ and $\boldsymbol{T}_{K_r,E(i)} : \widehat{S}_1 \to E(i)$. Since $\boldsymbol{S}_{lr}^{\mathrm{e}} := \boldsymbol{T}_{K_r,E(i)}^{-1} \circ \boldsymbol{T}_{K_l,E(i)} : \widehat{S}_1 \to \widehat{S}_1$ is an affine bijective mapping, and since we assumed that the basis $\{\mu_m\}_{m \in \{1:n_{\mathrm{sh}}^{\mathrm{e}}\}}$ is invariant under permutation of the vertices of $\widehat{S}^1$, there is an index permutation $\varpi_{lr}^{\mathrm{e}} : \{1{:}n_{\mathrm{sh}}^{\mathrm{e}}\} \to \{1{:}n_{\mathrm{sh}}^{\mathrm{e}}\}$ such that $\mu_m \circ \boldsymbol{S}_{lr}^{\mathrm{e}} = \mu_{\varpi_{lr}^{\mathrm{e}}(m)}$. By definition of $\mathcal{N}_{K_r,F}^{\mathrm{e}}$, there is $i_r \in \mathcal{N}_{K_r,F}^{\mathrm{e}}$ such that $E(i_r) = E(i)$ and $m(i_r) = \varpi_{lr}^{\mathrm{e}}(m(i))$.

Finally, we set $\chi_{lr}(i) := i_r$, and this gives $\sigma^{\mathrm{g}}_{K_r,F,\chi_{lr}(i)}(v) = \sigma^{\mathrm{g}}_{K_l,F,i}(v)$ for all $v \in P^{\mathrm{g}}_F$. Notice that $\chi_{lr}(\mathcal{N}^{\mathrm{e}}_{K_l,F}) = \mathcal{N}^{\mathrm{e}}_{K_r,F}$.
(iii.c) Assume $i \in \mathcal{N}^{\mathrm{s}}_{K_l,F}$ and let $m(i) \in \{1{:}n^{\mathrm{f}}_{\mathrm{sh}}\}$ be the associated index. Since $\boldsymbol{S}^{\mathrm{s}}_{lr} := \boldsymbol{T}^{-1}_{K_r,F} \circ \boldsymbol{T}_{K_l,F} : \widehat{S}^2 \to \widehat{S}^2$ is an affine bijective mapping, and since we assumed that the basis $\{\zeta_m\}_{m\in\{1:n^{\mathrm{f}}_{\mathrm{sh}}\}}$ is invariant under permutation of the vertices of $\widehat{S}^2$, there is an index permutation $\varpi^{\mathrm{s}}_{lr} : \{1{:}n^{\mathrm{f}}_{\mathrm{sh}}\} \to \{1{:}n^{\mathrm{f}}_{\mathrm{sh}}\}$ such that $\zeta_m \circ \boldsymbol{S}^{\mathrm{s}}_{lr} = \zeta_{\varpi^{\mathrm{s}}_{lr}(m)}$. Then by definition of $\mathcal{N}^{\mathrm{s}}_{K_r,F}$, there is $i_r \in \mathcal{N}^{\mathrm{s}}_{K_r,F}$ such that $m(i_r) = \varpi^{\mathrm{s}}_{lr}(m(i))$. Finally, we set $\chi_{lr}(i) := i_r$, and this gives $\sigma^{\mathrm{g}}_{K_r,F,\chi_{lr}(i)}(v) = \sigma^{\mathrm{g}}_{K_l,F,i}(v)$ for all $v \in P^{\mathrm{g}}_F$. Notice that $\chi_{lr}(\mathcal{N}^{\mathrm{s}}_{K_l,F}) = \mathcal{N}^{\mathrm{s}}_{K_r,F}$.
In conclusion, we have built a bijective map $\chi_{lr} : \mathcal{N}_{K_l,F} \to \mathcal{N}_{K_r,F}$ such that $\sigma^{\mathrm{g}}_{K_r,F,\chi_{lr}(i)} = \sigma^{\mathrm{g}}_{K_l,F,i}$ for all $i \in \mathcal{N}_{K_l,F}$.

Exercise 20.7 ($P_{K,F}$). (i) Let us set $\boldsymbol{x} := (x, y)$. Recall that $\widehat{\lambda}_0(\boldsymbol{x}) = 1 - x - y$, $\widehat{\lambda}_1(\boldsymbol{x}) = x$, $\widehat{\lambda}_1(\boldsymbol{x}) = y$, and that $\widehat{F}^0 = \{x + y = 1\}$, $\widehat{F}^1 = \{x = 0\}$, $\widehat{F}^2 = \{y = 0\}$. We have

$$\widehat{\lambda}_{0|\widehat{F}_0}(\boldsymbol{x}) = 0, \quad \widehat{\lambda}_{1|\widehat{F}_0}(\boldsymbol{x}) = x, \quad \widehat{\lambda}_{2|\widehat{F}_0}(\boldsymbol{x}) = 1 - x.$$

Hence, $\widehat{\phi}_{0,1}(\boldsymbol{x}) := x$, $\widehat{\phi}_{0,2}(\boldsymbol{x}) := 1 - x$ forms a basis of $P_{\widehat{K},\widehat{F}_0}$. Similarly, we have

$$\widehat{\lambda}_{0|\widehat{F}_1}(\boldsymbol{x}) = 1 - y, \quad \widehat{\lambda}_{1|\widehat{F}_1}(\boldsymbol{x}) = 0, \quad \widehat{\lambda}_{2|\widehat{F}_1}(\boldsymbol{x}) = y.$$

Hence, $\widehat{\phi}_{1,1}(\boldsymbol{x}) := y$, $\widehat{\phi}_{1,2}(\boldsymbol{x}) := 1 - y$ forms a basis of $P_{\widehat{K},\widehat{F}_1}$. Finally, we have

$$\widehat{\lambda}_{0|\widehat{F}_2}(\boldsymbol{x}) = 1 - x, \quad \widehat{\lambda}_{1|\widehat{F}_2}(\boldsymbol{x}) = x, \quad \widehat{\lambda}_{2|\widehat{F}_2}(\boldsymbol{x}) = 0.$$

Hence, $\widehat{\phi}_{2,1}(\boldsymbol{x}) := x$, $\widehat{\phi}_{2,2}(\boldsymbol{x}) := 1 - x$ forms a basis of $P_{\widehat{K},\widehat{F}_2}$. For $\widehat{F}_0$, we have $\Sigma_{\widehat{K},\widehat{F}_0} = \{\widehat{\sigma}_{0,1}, \widehat{\sigma}_{0,2}\}$ with $\widehat{\sigma}_{0,2}(\widehat{p}) := \widehat{p}(\boldsymbol{z}_1)$, $\widehat{\sigma}_{0,2}(\widehat{p}) := \widehat{p}(\boldsymbol{z}_2)$ for all $\widehat{p} \in P_{\widehat{K},\widehat{F}_0}$, and the triple $(\widehat{F}_0, P_{\widehat{K},\widehat{F}_0}, \Sigma_{\widehat{K},\widehat{F}_0})$ is a finite element. The reasoning for $\widehat{F}_1$ and $\widehat{F}_2$ is similar.
(ii) Since the polynomial space for the Crouzeix–Raviart element is the same as for the $\mathbb{P}_1$ Lagrange element, the bases are those found in Step (i). Since for the Crouzeix–Raviart element, there is only one dof per face, we have $\mathrm{card}(\Sigma_{\widehat{K},\widehat{F}_i}) = 1$. For instance, we have

$$\sigma_{\widehat{K},\widehat{F}_0}(\widehat{p}) = \widehat{p}\big(\tfrac{1}{2}(\boldsymbol{z}_1 + \boldsymbol{z}_2)\big),$$

for all $\widehat{p} \in P_{\widehat{K},\widehat{F}_0}$. In conclusion, $(\widehat{F}_i, P_{\widehat{K},\widehat{F}_i}, \Sigma_{\widehat{K},\widehat{F}_i})$ is not a finite element since $\dim(P_{\widehat{K},\widehat{F}_i}) = 2$ and $\mathrm{card}(\Sigma_{\widehat{K},\widehat{F}_i}) = 1$.

Chapter 21

Construction of the connectivity classes

Exercises

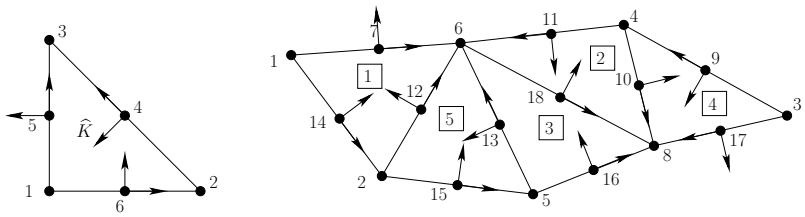
Exercise 21.1 (Mesh orientation, $\mathcal{N}_{K,F}$, χ_{lr}). Consider the mesh $\mathcal{T}_h$ shown in Exercise 19.1. (i) Orient the mesh by using the increasing vertex-index enumeration technique. (ii) Consider the corresponding space $P_2^{\mathrm{g}}(\mathcal{T}_h)$. Use the enumeration convention adopted in this chapter for the dofs. Find the two cells K_l, K_r for the second face of the cell 5 and for the first face of the cell 3. (iii) Let F be the second face of the cell 5. Identify $\mathcal{N}_{5,F}$, `j_dof`$(5, \mathcal{N}_{5,F})$, and the map χ_{lr}. (iv) Let F' be the first face of the cell 3. Identify $\mathcal{N}_{3,F'}$, `j_dof`$(3, \mathcal{N}_{3,F'})$, and the map χ_{lr}.

Exercise 21.2 (M-dofs). Let $K \in \mathcal{T}_h$, let $F \in \mathcal{F}_K$, and let $M \in \mathcal{M}_h$ be a geometric entity s.t. $M \subset F$. Prove that $\mathcal{N}_{K,M} \subset \mathcal{N}_{K,F}$.

Exercise 21.3 ($\mathbb{Q}_{k,3}$ dofs). Determine $n_{\mathrm{sh}}^{\mathrm{v}}, n_{\mathrm{sh}}^{\mathrm{e}}, n_{\mathrm{sh}}^{\mathrm{f}}, n_{\mathrm{sh}}^{\mathrm{c}}$ for scalar-valued $\mathbb{Q}_{k,3}$ Lagrange elements.

Solution to exercises

Exercise 21.1 (Mesh orientation, $\mathcal{N}_{K,F}$, χ_{lr}). (i) Here is a picture showing the orientation vectors $\{\boldsymbol{\tau}_E\}_{E\in\mathcal{E}_h}$, $\{\boldsymbol{n}_F\}_{F\in\mathcal{F}_h}$ for the mesh in question:



(ii) Recalling the orientation convention of the faces, we have $K_l = 5$ and $K_r = 1$ for the second face of the cell 5 (this is the face of K_5 whose vertices have indices 2 and 6), and we have $K_l = 3$ and $K_r = 2$ or the first face of the cell 3 (this is the face of K_3 whose vertices have indices 6 and

8).
(iii) For the interface F, we have

$$\begin{aligned}\mathcal{N}_{5,F} &= \{1,3,5\},\\ \texttt{j\_dof}(5,\mathcal{N}_{5,F}) &= \{2,6,12\}.\end{aligned}$$

To find the map χ_{lr}, we first identify

$$\begin{aligned}\mathcal{N}_{1,F} &= \{2,3,4\},\\ \texttt{j\_dof}(1,\mathcal{N}_{1,F}) &= \{2,6,12\},\end{aligned}$$

so that

$$\chi_{lr}(1) = 2, \quad \chi_{lr}(3) = 3, \quad \chi_{lr}(5) = 4.$$

(iv) For the interface F', we have

$$\begin{aligned}\mathcal{N}_{3,F'} &= \{2,3,4\},\\ \texttt{j\_dof}(3,\mathcal{N}_{3,F'}) &= \{6,8,18\}.\end{aligned}$$

To find the map χ_{lr}, we first identify

$$\begin{aligned}\mathcal{N}_{2,F'} &= \{2,3,4\},\\ \texttt{j\_dof}(1,\mathcal{N}_{2,F'}) &= \{6,8,18\},\end{aligned}$$

so that

$$\chi_{lr}(2) = 2, \quad \chi_{lr}(3) = 3, \quad \chi_{lr}(4) = 4.$$

Exercise 21.2 (M-dofs). Let $i \in \mathcal{N}\backslash\mathcal{N}_{K,F}$, i.e., $i \notin \mathcal{N}_{K,F}$. Assumption 20.12 implies that $\theta_{K,i} \in \ker(\gamma_{K,F})$. Assume that $i \in \mathcal{N}_{K,M}$. Then $\sigma_{K,i} = \sigma_{K,M,i} \circ \gamma_{K,M}$ and $\ker(\gamma_{K,F}) \subset \ker(\gamma_{K,M})$ by Assumption 21.9. The inclusion $\ker(\gamma_{K,F}) \subset \ker(\gamma_{K,M})$ means that $\sigma_{K,i}(\theta_{K,i}) = \sigma_{K,M,i} \circ \gamma_{K,M}(\theta_{K,i}) = 0$, which is absurd. Hence, $i \notin \mathcal{N}_{K,M}$. This proves that $\mathcal{N}_{K,M} \subset \mathcal{N}_{K,F}$.

Exercise 21.3 ($\mathbb{Q}_{k,3}$ dofs). We have $n_{\mathrm{sh}}^{\mathrm{v}} = 1$, $n_{\mathrm{sh}}^{\mathrm{e}} = k-1$ if $k \geq 2$, $n_{\mathrm{sh}}^{\mathrm{f}} = (k-1)^2$ if $k \geq 2$, and $n_{\mathrm{sh}}^{\mathrm{c}} = (k-1)^3$ if $k \geq 2$.

Chapter 22

Quasi-interpolation and best approximation

Exercises

Exercise 22.1 ($\check{\mathcal{F}}_K^\circ$). Identify the set $\check{\mathcal{F}}_K^\circ$ for the canonical hybrid, Nédélec, and Raviart–Thomas elements.

Exercise 22.2 (L^p-stability). Prove directly, i.e., without using Lemma 22.3, the L^p-stability of $\mathcal{J}_h^{\mathrm{av}}$. (*Hint*: use Proposition 12.5.)

Exercise 22.3 (Poincaré–Steklov in D_K). The goal is to prove (22.20). Let $p \in [1,\infty]$, $K \in \mathcal{T}_h$, and $v \in W^{1,p}(D_K)$ (i) Let $K_l, K_r \in \check{\mathcal{T}}_K$ sharing an interface $F := \partial K_l \cap \partial K_r$. Show that

$$|K|^{\frac{1}{p}} |\underline{v}_{K_l} - \underline{v}_{K_r}| \le c\, h_K |v|_{W^{1,p}(K_l \cup K_r)}.$$

(*Hint*: observe that $|F|^{-\frac{1}{p}} |\underline{v}_{K_l} - \underline{v}_{K_r}| \le \|v_{K_l} - \underline{v}_{K_l}\|_{L^p(F)} + \|v_{K_r} - \underline{v}_{K_r}\|_{L^p(F)}$, then use the trace inequality (12.16).) (ii) Prove (22.20). (*Hint*: use that $\underline{v}_{D_K} - \underline{v}_{K'} = \sum_{K'' \in \check{\mathcal{T}}_K} \frac{|K''|}{|D_K|} (\underline{v}_{K''} - \underline{v}_{K'})$ for all $K' \in \check{\mathcal{T}}_K$.)

Exercise 22.4 (Polynomial approximation in D_K). Prove that there is c s.t. for all $r \in [0, k+1]$, all $p \in [1,\infty)$ if $r \notin \mathbb{N}$ or all $p \in [1,\infty]$ if $r \in \mathbb{N}$, every integer $m \in \{0 : \lfloor r \rfloor\}$, all $v \in W^{r,p}(D_K)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$:

$$\inf_{g \in \mathbb{P}_{k,d}} |v - g|_{W^{m,p}(D_K)} \le c\, h_K^{r-m} |v|_{W^{r,p}(D_K)}. \tag{22.1}$$

(*Hint*: use Morrey's polynomial as in the proof of Corollary 12.13.)

Exercise 22.5 (Approximation on faces). (i) Prove that

$$\|v - \mathcal{I}_h^{\mathrm{av}}(v)\|_{L^p(F)} \le c\, h_K^{r-\frac{1}{p}} |v|_{W^{r,p}(\check{\mathcal{T}}_K)},$$

for all $p \in [1,\infty)$, all $r \in (\frac{1}{p}, k+1]$ if $p > 1$ or $r \in [1, k+1]$ if $p = 1$, all $v \in W^{r,p}(D_K)$, all $K \in \mathcal{T}_h$, all $F \in \mathcal{F}_K$, and all $h \in \mathcal{H}$ (c can grow unboundedly as $rp \downarrow 1$ if $p > 1$). (*Hint*: use the

multiplicative trace inequality (12.16) or its fractional version (12.17).) (ii) Assume $k \geq 1$. Prove that

$$\|\nabla(v - \mathcal{I}_h^{\mathrm{av}}(v))\|_{\boldsymbol{L}^p(F)} \leq c\, h_K^{r-\frac{1}{p}} |v|_{W^{1+r,p}(\check{\mathcal{T}}_K)},$$

for all $r \in (\frac{1}{p}, k]$ if $p > 1$ or $r \in [1, k]$ if $p = 1$, all $v \in W^{1+r,p}(D_K)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$.

Exercise 22.6 (L^2-projection). (i) Prove that (22.42) implies the H^1-stability of $\mathcal{P}_h^{\mathrm{g}}$. (*Hint*: adapt the proof of Proposition 22.21.) (ii) Set $\|y\|_{*,r} := \sup_{w \in H^r(D;\mathbb{R}^q)} \frac{(y,w)_{L^2(D;\mathbb{R}^q)}}{\|w\|_{H^r(D;\mathbb{R}^q)}}$ for all $y \in L^2(D;\mathbb{R}^q)$ (this is not the standard norm of the dual space $H^{-r}(D;\mathbb{R}^q) := (H_0^r(D;\mathbb{R}^q))'$). Prove that there is c s.t. for every integer $r \in \{1{:}k+1\}$, all $v \in L^2(D;\mathbb{R}^q)$, and all $h \in \mathcal{H}$,

$$\|v - \mathcal{P}_h(v)\|_{*,r} \leq c\, h^r \|v - \mathcal{P}_h(v)\|_{L^2(D;\mathbb{R}^q)},$$
$$\|v - \mathcal{P}_{h0}(v)\|_{H^{-r}(D;\mathbb{R}^q)} \leq c\, h^r \|v - \mathcal{P}_{h0}(v)\|_{L^2(D;\mathbb{R}^q)}.$$

(*Hint*: use $\mathcal{I}_h^{\mathrm{av}}(v)$.)

Exercise 22.7 (Discrete commutator). Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular mesh sequence. The goal is to prove that there is c s.t. for every integers $l \in \{0{:}1\}$ and $m \in \{0{:}l\}$, all $p \in [1,\infty]$, all $v_h \in P_k^{\mathrm{g}}(\mathcal{T}_h)$, all $K \in \mathcal{T}_h$, all $h \in \mathcal{H}$, and all ϕ in $W^{1+l,\infty}(D)$,

$$\|\phi v_h - \mathcal{I}_h^{\mathrm{g,av}}(\phi v_h)\|_{W^{m,p}(K)} \leq c\, h_K^{1+l-m} \|v_h\|_{W^{l,p}(D_K)} \|\phi\|_{W^{1+l,\infty}(D_K)}.$$

This property provides a useful tool to analyze nonlinear problems; see Bertoluzza [4] and Johnson and Szepessy [31]. (i) Fix $K \in \mathcal{T}_h$. Let $\underline{v}_{D_K}$ denote the mean value of v_h in D_K. Prove that

$$\|\phi \underline{v}_{D_K} - \mathcal{I}_h^{\mathrm{g,av}}(\phi \underline{v}_{D_K})\|_{W^{m,p}(K)} \leq c\, h_K^{1+l-m} \|v_h\|_{L^p(D_K)} \|\phi\|_{W^{1+l,\infty}(D_K)}.$$

(*Hint*: use Theorem 22.6 and verify that $\|\underline{v}_{D_K}\|_{L^p(D_K)} \leq \|v_h\|_{L^p(D_K)}$.) (ii) Set $\eta_h := v_h - \underline{v}_{D_K}$. Prove that

$$\|\phi\eta_h - \mathcal{I}_h^{\mathrm{g,av}}(\phi\eta_h)\|_{W^{m,p}(K)} \leq c h_K^{1+l-m} \|v_h\|_{W^{l,p}(D_K)} \|\phi\|_{W^{1,\infty}(D_K)}.$$

(*Hint*: observe that $\phi(\boldsymbol{x}_K)\eta_h = \mathcal{I}_h^{\mathrm{g,av}}(\phi(\boldsymbol{x}_K)\eta_h)$ where $\boldsymbol{x}_K$ is some point in K, e.g., the barycenter of K, then use (22.20) to bound η_h.) Conclude.

Solution to exercises

Exercise 22.1 ($\check{\mathcal{F}}_K^\circ$). For the canonical hybrid element, the set $\check{\mathcal{F}}_K^\circ$ collects all the mesh interfaces that have at least a common vertex with K. For Nédélec elements, the set $\check{\mathcal{F}}_K^\circ$ collects all the mesh interfaces that have at least a common edge with K. For Raviart–Thomas elements, the set $\check{\mathcal{F}}_K^\circ$ collects all the mesh interfaces that are faces of K.

Exercise 22.2 (L^p-stability). We prove the result for $p = \infty$. The other cases are obtained by using local inverse inequalities in $P^{\mathrm{b}}(\mathcal{T}_h)$. Using the triangle inequality and the regularity of the

mesh sequence, we infer that

$$\begin{aligned}\|\mathcal{J}_h^{\mathrm{av}}(v_h)\|_{L^\infty(K;\mathbb{R}^q)} &\le \sum_{i\in\mathcal{N}} \frac{\|\theta_{K,i}\|_{L^\infty(K;\mathbb{R}^q)}}{\operatorname{card}(a_{K,i})} \sum_{(K',i')\in a_{K,i}} \big|\sigma_{K',i'}(v_{h|K'})\big| \\ &\le c\sum_{i\in\mathcal{N}} \frac{\|\mathbb{A}_K^{-1}\|_{\ell^2}}{\operatorname{card}(a_{K,i})} \sum_{(K',i')\in a_{K,i}} \big|\sigma_{K',i'}(v_{h|K'})\big| \\ &\le c\sum_{K'\in\mathcal{T}_K} \|\mathbb{A}_K^{-1}\|_{\ell^2} \sum_{i'\in\mathcal{N}} \big|\sigma_{K',i'}(v_{h|K'})\big| \\ &\le c\,\|v_h\|_{L^\infty(D_K;\mathbb{R}^q)},\end{aligned}$$

where we used $|\sigma_{K',i'}(v_{h|K'})| \le |\sigma_{K',i'}(v_{h|K'}) - \sigma_{K,i}(v_{h|K})| + |\sigma_{K,i}(v_{h|K})|$, the assumption (22.8), the inequality $|[\![v_h]\!]_F|_{L^\infty(F;\mathbb{R}^t)} \le \|v_h\|_{L^\infty(D_K;\mathbb{R}^q)}$ and Proposition 12.5.

Exercise 22.3 (Poincaré–Steklov in D_K). (i) To prove the hint, we observe that

$$\begin{aligned}|\underline{v}_{K_l} - \underline{v}_{K_r}| &= |F|^{-\frac1p}\|\underline{v}_{K_l} - \underline{v}_{K_r}\|_{L^p(F)} \\ &= |F|^{-\frac1p}\|\underline{v}_{K_l} - v_{|K_l} + v_{|K_r} - \underline{v}_{K_r}\|_{L^p(F)},\end{aligned}$$

since $v_{|K_l} = v_{|K_r}$ in F owing to Theorem 18.8, and we conclude by using the triangle inequality. We can now bound each of norms $\|v_{|K_i} - \underline{v}_{K_i}\|_{L^p(F)}$, $i\in\{l,r\}$, using the trace inequality (12.17) (see also Exercise 12.6) and the fact that $\|v - \underline{v}_{K_i}\|_{L^p(K_i)} \le \frac{1}{\pi} h_{K_i}|v|_{W^{1,p}(K_i)}$ (owing to (12.13) since K_i is a convex set). This yields $\|v_{|K_i} - \underline{v}_{K_i}\|_{L^p(F)} \le c h_{K_i}^{1-\frac1p}|v|_{W^{1,p}(K_i)}$, and we conclude by invoking the regularity of the mesh sequence.
(ii) Let $K'\in\check{\mathcal{T}}_K$. Using the hint and the triangle inequality, we observe that

$$\|v - \underline{v}_{D_K}\|_{L^p(K')} \le \|v - \underline{v}_{K'}\|_{L^p(K')} + \sum_{K''\in\check{\mathcal{T}}_K} \frac{|K''|}{|D_K|}|\underline{v}_{K''} - \underline{v}_{K'}||K'|^{\frac1p}.$$

For all $K''\in\check{\mathcal{T}}_K$, we can find a path of mesh cells in $\check{\mathcal{T}}_K$ linking K' to K'' s.t. any consecutive mesh cells in the path share a common face. Using Step (i) together with the regularity of the mesh sequence, we infer that $\|v - \underline{v}_{K'}\|_{L^p(K')} \le c h_K |v|_{W^{1,p}(D_K)}$, and the conclusion follows by summing over $K'\in\mathcal{T}_h$ and using the fact that $\operatorname{card}(\check{\mathcal{T}}_K)$ is uniformly bounded.

Exercise 22.4 (Polynomial approximation in D_K). We proceed as in Bramble and Hilbert [5, Thm. 1], but instead of invoking Morrey [34, Thm. 3.6.11], where the constants may depend on D_K, we are going to track the constants to make sure that they are independent of D_K. If $m = r$, there is nothing to prove. Let us assume that $m < r$. Let $\ell\in\mathbb{N}$ be such that $\ell = r-1$ if r is a natural number or $\ell = \lfloor r\rfloor$ otherwise (note that $1\le r$ if r is a natural number since we assumed that $0\le m<r$). In both cases, the integer ℓ is such that $m\le\ell\le k$. Let $\mathcal{A}_{\ell,d} = \{\alpha\in\mathbb{N}^d \mid |\alpha| := \alpha_1+\ldots+\alpha_d \le \ell\}$. Note that $\operatorname{card}(\mathcal{A}_{\ell,d}) = \dim(\mathbb{P}_{\ell,d}) = \binom{\ell+d}{d} =: N_{\ell,d}$. Since the map $\Phi_{\ell,d}:\mathbb{P}_{\ell,d}\to\mathbb{R}^{N_{\ell,d}}$ such that $\Phi_{\ell,d}(q) = (\int_{D_K}\partial^\alpha q\,\mathrm{d}x)_{\alpha\in\mathcal{A}_{\ell,d}}$ is an isomorphism, there is a unique polynomial $\pi_\ell(v)\in\mathbb{P}_{\ell,d}$ such that $\Phi_{\ell,d}(\pi_\ell(v)) = (\int_{D_K}\partial^\alpha v\,\mathrm{d}x)_{\alpha\in\mathcal{A}_{\ell,d}}$, i.e., $\int_{D_K}\partial^\alpha(v-\pi_\ell(v))\,\mathrm{d}x = 0$ for all $\alpha\in\mathcal{A}_{\ell,d}$ (this result is actually stated in Morrey [34, Thm. 3.6.10]).

Since by definition $\int_{D_K}\partial^\alpha(v-\pi_\ell(v))\,\mathrm{d}x = 0$ for all $|\alpha| = m\le\ell$, we can apply (22.20), i.e., there is a uniform constant c such that $|v-\pi_\ell(v)|_{W^{m,p}(D_K)} \le c h_K |v-\pi_\ell(v)|_{W^{m+1,p}(D_K)}$. We can repeat the argument if $m+1\le\ell$ since in this case we also have $\int_{D_K}\partial^\alpha(v-\pi_\ell(v))\,\mathrm{d}x = 0$ for all $|\alpha| = m+1\le\ell$. Eventually, we obtain

$$|v-\pi_\ell(v)|_{W^{m,p}(D_K)} \le c h_K^{\ell-m}|v-\pi_\ell(v)|_{W^{\ell,p}(D_K)}.$$

If $r \in \mathbb{N}$, then $\ell + 1 = r$, and we can apply the above argument one last time since $\int_{D_K} \partial^\alpha(v - \pi_\ell(v))\,\mathrm{d}x = 0$ for all $|\alpha| = \ell$, which gives (22.1) because $\partial^\alpha \pi_\ell(v) = 0$ for all $|\alpha| = \ell+1$. Otherwise, $\ell = \lfloor r \rfloor$ and we apply Lemma 3.26 to all the partial derivatives $\partial^\alpha(v - \pi_\ell(v))$ with $|\alpha| = \ell$, $s = r - \lfloor r \rfloor \in (0,1)$ and $O := D_K$; this is legitimate since all these partial derivatives have zero average over $O = D_K$. We infer that there is c, uniform with respect to s, p, K, and v, such that

$$|v - \pi_\ell(v)|_{W^{m,p}(D_K;\mathbb{R}^q)} \le c\, h_K^{\lfloor r \rfloor - m} h_{D_K}^{r - \lfloor r \rfloor} \left(\frac{h_{D_K}^d}{|D_K|}\right)^{\frac{1}{p}} |v - \pi_\ell(v)|_{W^{r,p}(D_K;\mathbb{R}^q)}.$$

Note that $|v - \pi_\ell(v)|_{W^{r,p}(D_K;\mathbb{R}^q)} = |v|_{W^{r,p}(D_K;\mathbb{R}^q)}$ since $\partial^\alpha \pi_\ell(v)$ is a constant in $\mathbb{R}^q$ for all $|\alpha| = \ell$. We conclude that (22.1) holds true owing to the regularity of the mesh sequence.

Exercise 22.5 (Approximation on faces). (i) Let us assume first that $r \in [1, k+1]$. We can invoke the multiplicative trace inequality (12.16). Letting $\eta := v - \mathcal{I}_h^{\mathrm{av}}(v)$, we have

$$\|\eta\|_{L^p(F)} \le c\Big(h_K^{-\frac{1}{p}} \|\eta\|_{L^p(K)} + \|\eta\|_{L^p(K)}^{1-\frac{1}{p}} |\eta|_{W^{1,p}(K)}^{\frac{1}{p}}\Big).$$

The expected bound follows from Theorem 22.6 (with $m \in \{0,1\}$). Let us now assume that $r \in (\frac{1}{p}, 1)$ with $p > 1$. Let $\underline{v}$ be the mean value of v in K. We have

$$\begin{aligned}
h_K^{\frac{1}{p}} \|v - \mathcal{I}_h^{\mathrm{av}}(v)\|_{L^p(F)} &\le h_K^{\frac{1}{p}} \|v - \underline{v}\|_{L^p(F)} + h_K^{\frac{1}{p}} \|\underline{v} - \mathcal{I}_h^{\mathrm{av}}(v)\|_{L^p(F)} \\
&\le c(\|v - \underline{v}\|_{L^p(K)} + h_K^r |v|_{W^{r,p}(K)}) + h_K^{\frac{1}{p}} \|\underline{v} - \mathcal{I}_h^{\mathrm{av}}(v)\|_{L^p(F)} \\
&\le c'(\|v - \underline{v}\|_{L^p(K)} + \|\underline{v} - \mathcal{I}_h^{\mathrm{av}}(v)\|_{L^p(K)} + h_K^r |v|_{W^{r,p}(K)}) \\
&\le c'(2\|v - \underline{v}\|_{L^p(K)} + \|v - \mathcal{I}_h^{\mathrm{av}}(v)\|_{L^p(K)} + h_K^r |v|_{W^{r,p}(K)}),
\end{aligned}$$

where we used the triangle inequality in the first line, (12.16) if $r = 1$ or (12.17) if $r < 1$ and the fact that $|v - \underline{v}|_{W^{r,p}(K)} = |v|_{W^{r,p}(K)}$ since $\underline{v}$ is constant on K in the second line, the discrete trace inequality (12.10) (with $r := p$) in the third line, and the triangle inequality in the fourth line. We conclude by using the Poincaré–Steklov inequality (12.14) and the approximation properties of $\mathcal{I}_h^{\mathrm{av}}$ from Theorem 22.6 (with $m := 0$).
(ii) We proceed as above. If $r \in [1, k]$, the desired bound follows from the multiplicative trace inequality (12.16) and Theorem 22.6 (with $m \in \{1,2\}$). Otherwise, let us assume $r \in [\frac{1}{p}, 1)$ and $p > 1$. Let $\underline{\boldsymbol{w}}$ be the average over K of ∇v. We have

$$\begin{aligned}
h_K^{\frac{1}{p}} \|\nabla v - \nabla \mathcal{I}_h^{\mathrm{av}}(v)\|_{\boldsymbol{L}^p(F)} &\le h_K^{\frac{1}{p}} \|\nabla v - \underline{\boldsymbol{w}}\|_{\boldsymbol{L}^p(F)} + h_K^{\frac{1}{p}} \|\underline{\boldsymbol{w}} - \nabla \mathcal{I}_h^{\mathrm{av}}(v)\|_{\boldsymbol{L}^p(F)} \\
&\le c(\|\nabla v - \underline{\boldsymbol{w}}\|_{\boldsymbol{L}^p(K)} + h_K^r |v|_{W^{r+1,p}(K)}) + h_K^{\frac{1}{p}} \|\underline{\boldsymbol{w}} - \nabla \mathcal{I}_h^{\mathrm{av}}(v)\|_{\boldsymbol{L}^p(F)} \\
&\le c'(\|\nabla v - \underline{\boldsymbol{w}}\|_{\boldsymbol{L}^p(K)} + \|\underline{\boldsymbol{w}} - \nabla \mathcal{I}_h^{\mathrm{av}}(v)\|_{\boldsymbol{L}^p(K)} + h_K^r |v|_{W^{r+1,p}(K)}) \\
&\le c'(2\|\nabla v - \underline{\boldsymbol{w}}\|_{\boldsymbol{L}^p(K)} + \|\nabla v - \nabla \mathcal{I}_h^{\mathrm{av}}(v)\|_{\boldsymbol{L}^p(K)} + h_K^r |v|_{W^{r+1,p}(K)}).
\end{aligned}$$

We conclude using the Poincaré–Steklov inequality (12.14) (componentwise) and the approximation properties of $\mathcal{I}_h^{\mathrm{av}}$ from Theorem 22.6 (with $m := 1$).

Exercise 22.6 (L^2-projection). (i) We employ the same arguments as in the proof of Proposition 22.21, except that we use a local inverse inequality and the $\hbar^{-1}$-weighted stability property (22.42). This yields

$$\begin{aligned}
|\mathcal{P}_h^{\mathrm{g}}(v)|_{H^1(D)} &\le |\mathcal{P}_h^{\mathrm{g}}(v - \mathcal{I}_h^{\mathrm{g,av}}(v))|_{H^1(D)} + |\mathcal{I}_h^{\mathrm{g,av}}(v)|_{H^1(D)} \\
&\le c\,\|\hbar^{-1} \mathcal{P}_h^{\mathrm{g}}(v - \mathcal{I}_h^{\mathrm{g,av}}(v))\|_{L^2(D)} + |\mathcal{I}_h^{\mathrm{g,av}}(v)|_{H^1(D)} \\
&\le c\,\|\hbar^{-1}(v - \mathcal{I}_h^{\mathrm{g,av}}(v))\|_{L^2(D)} + |\mathcal{I}_h^{\mathrm{g,av}}(v)|_{H^1(D)},
\end{aligned}$$

and we conclude by invoking Theorem 22.6.
(ii) For all $v \in L^2(D;\mathbb{R}^q)$ and all $w \in H^r(D;\mathbb{R}^q)$, we observe that

$$\begin{aligned}(v-\mathcal{P}_h(v),w)_{L^2(D;\mathbb{R}^q)} &= (v-\mathcal{P}_h(v),w-\mathcal{I}_h^{\mathrm{av}}(w))_{L^2(D;\mathbb{R}^q)}\\ &\le \|v-\mathcal{P}_h(v)\|_{L^2(D;\mathbb{R}^q)}\|w-\mathcal{I}_h^{\mathrm{av}}(w)\|_{L^2(D;\mathbb{R}^q)}\\ &\le \|v-\mathcal{P}_h(v)\|_{L^2(D;\mathbb{R}^q)}ch^r|w|_{H^r(D;\mathbb{R}^q)},\end{aligned}$$

where we used (22.35), the Cauchy–Schwarz inequality, and Theorem 22.6 (with $p := 2$ and $m := 0$). The proof of the second inequality is almost identical since one has to invoke $\mathcal{I}_{h0}^{\mathrm{av}}(w)$ instead of $\mathcal{I}_h^{\mathrm{av}}(w)$ (and Theorem 22.14).

Exercise 22.7 (Discrete commutator). (i) Owing to Theorem 22.6 (since $m \le 1+l$), we infer that

$$\begin{aligned}\|\phi\underline{v}_{D_K}-\mathcal{I}_h^{\mathrm{g,av}}(\phi\underline{v}_{D_K})\|_{W^{m,p}(K)} &\le c\,h_K^{1+l-m}|\phi\underline{v}_{D_K}|_{W^{1+l,p}(D_K)}\\ &= c\,h_K^{1+l-m}\|\underline{v}_{D_K}\|_{L^p(D_K)}\|\phi\|_{W^{1+l,\infty}(D_K)},\end{aligned}$$

since $\underline{v}_{D_K}$ is constant. We conclude by observing that $\|\underline{v}_{D_K}\|_{L^p(D_K)} \le \|v_h\|_{L^p(D_K)}$ since, owing to Hölder's inequality with $p' := \frac{p}{p-1}$, we have

$$\underline{v}_{D_K} = |D_K|^{-1}\int_{D_K} v_h\,\mathrm{d}x \le |D_K|^{-1}\|v_h\|_{L^p(D_K)}\|1\|_{L^{p'}(D_K)} = |D_K|^{-\frac{1}{p}}\|v_h\|_{L^p(D_K)}.$$

(ii) The hint is proved by observing that $\phi(\boldsymbol{x}_K)$ is constant and that $P_k^{\mathrm{g}}(\mathcal{T}_h)$ is pointwise invariant under $\mathcal{I}_h^{\mathrm{g,av}}$. As a result, letting $\eta_\phi := \phi - \phi(\boldsymbol{x}_K)$, we infer that

$$\begin{aligned}\|\phi\eta_h-\mathcal{I}_h^{\mathrm{g,av}}(\phi\eta_h)\|_{W^{m,p}(K)} &= \|\eta_\phi\eta_h-\mathcal{I}_h^{\mathrm{g,av}}(\eta_\phi\eta_h)\|_{W^{m,p}(K)}\\ &\le c\,h_K^{1-m}\|\eta_\phi\eta_h\|_{W^{1,p}(D_K)}\\ &\le c\,h_K^{1-m}\Big(\|\eta_\phi\|_{L^\infty(D_K)}\|\eta_h\|_{W^{1,p}(D_K)}+|\eta_\phi|_{W^{1,\infty}(D_K)}\|\eta_h\|_{L^p(D_K)}\Big),\end{aligned}$$

where we used Theorem 22.6 (since $m \le 1$) followed by the Leibniz product rule. One readily verifies that $|\eta_\phi|_{W^{1,\infty}(D_K)} = |\phi|_{W^{1,\infty}(D_K)}$ and that $\|\eta_\phi\|_{L^\infty(D_K)} \le h_K|\phi|_{W^{1,\infty}(D_K)}$ owing to the fundamental theorem of calculus. Moreover, $\|\eta_h\|_{L^p(D_K)} \le ch_K|v_h|_{W^{1,p}(D_K)}$ owing to (22.20), and we have $|\eta_h|_{W^{1,p}(D_K)} = |v_h|_{W^{1,p}(D_K)}$. This yields the expected bound on $\|\phi\eta_h-\mathcal{I}_h^{\mathrm{g,av}}(\phi\eta_h)\|_{W^{m,p}(K)}$. Summing the two bounds and using the triangle inequality yields the assertion.

Chapter 23

Commuting quasi-interpolation

Exercises

Exercise 23.1 (Star-shaped domain). Assume that $\mathbf{0} \in D$ and that D is star-shaped with respect to the ball $B(\mathbf{0}, r)$ for some $r > 0$. Verify that the mapping $\boldsymbol{\varphi}_\delta : \mathbb{R}^d \to \mathbb{R}^d$ such that $\boldsymbol{\varphi}(\boldsymbol{x}) := (1-\delta)\boldsymbol{x}$ verifies the properties stated in Lemma 23.1.

Exercise 23.2 (Commuting). Prove Lemma 23.3. (*Hint*: use Lemma 9.6.)

Exercise 23.3 (Translation). Let $\lambda_0 > 0$. Assume that $\boldsymbol{\psi}_\lambda : D \to D$ is a diffeomorphism of class C^1 such that $\|\boldsymbol{\psi}_\lambda(\boldsymbol{x}) - \boldsymbol{x}\|_{\ell^2} \le c'\lambda$ and $\|D\boldsymbol{\psi}_\lambda(\boldsymbol{x}) - \mathbb{I}\|_{\ell^2} \le \frac{1}{2}$ for all $\boldsymbol{x} \in D$ and all $\lambda \in [0, \lambda_0]$. Assume also that $\boldsymbol{\mu}_{\lambda,t} : \boldsymbol{x} \mapsto \boldsymbol{x} + t(\boldsymbol{\psi}_\lambda(\boldsymbol{x}) - \boldsymbol{x})$ maps D into D for all $t \in [0,1]$ and all $\lambda \in [0, \lambda_0]$. Show that there is c such that $\|f \circ \boldsymbol{\psi}_\lambda - f\|_{L^p(D)} \le c\,\lambda \|\nabla f\|_{\boldsymbol{L}^p(D)}$ for all $\lambda \in [0, \lambda_0]$, all $f \in W^{1,p}(D)$, and all $p \in [1, \infty]$. (*Hint*: assume first that f is smooth, then use Remark 23.8.)

Exercise 23.4 (Approximation). (i) Prove (23.9) for $\mathcal{K}_\delta^{\mathrm{g}}$ with $s \in (0,1)$, $p \in [1, \infty)$. (ii) Prove the result for $s = 1$, $p \in [1, \infty]$. (*Hint*: use Exercise 23.3.) (iii) Prove (23.9) for $\mathcal{K}_\delta^{\mathrm{x}}$ for $\mathrm{x} \in \{\mathrm{c}, \mathrm{d}, \mathrm{b}\}$. (*Hint*: observe that $\mathcal{K}_\delta^{\mathrm{x}}(f) = \mathbb{K}^{\mathrm{x}} \mathcal{K}_\delta^{\mathrm{g}}(f)$.)

Exercise 23.5 (Preserving constants). Propose a definition of $\mathcal{K}_\delta$ that preserves constants and commutes with the differential operators. (*Hint*: start with $\check{\mathcal{K}}_\delta^{\mathrm{g}}(f) := \mathcal{K}_\delta^{\mathrm{g}}(f - \underline{f} - \underline{\nabla f}\cdot(\boldsymbol{x} - \boldsymbol{x}_D)) + \underline{f} + \underline{\nabla f}\cdot(\boldsymbol{x} - \boldsymbol{x}_D)$, $\underline{f}, \underline{\nabla f}$ denoting mean values over D and $\boldsymbol{x}_D$ the barycenter of D.)

Exercise 23.6 (Inverse inequality). Prove (23.19). (*Hint*: use (23.15b).)

Exercise 23.7 (Approximation with $\mathcal{J}_h^{\mathrm{c}}$). Let $r \in [0, k+1]$ and $p \in [1, \infty]$. Let $\boldsymbol{g} \in \boldsymbol{W}^{r,p}(D)$ be such that $\nabla \times \boldsymbol{g} \in \boldsymbol{W}^{r,p}(D)$. Prove that $\|\boldsymbol{g} - \mathcal{J}_h^{\mathrm{c}}(\boldsymbol{g})\|_{\boldsymbol{L}^p(D)} \le c h^r |\boldsymbol{g}|_{\boldsymbol{W}^{r,p}(D)}$ and $\|\nabla \times (\boldsymbol{g} - \mathcal{J}_h^{\mathrm{c}}(\boldsymbol{g}))\|_{\boldsymbol{L}^p(D)} \le c h^r |\nabla \times \boldsymbol{g}|_{\boldsymbol{W}^{r,p}(D)}$. (*Hint*: use Theorem 23.12.)

Exercise 23.8 (Best approximation in L^p). We propose an alternative proof of Corollary 22.9 on quasi-uniform meshes. Let $h \in \mathcal{H}$ be the meshsize of $\mathcal{T}_h$ and set $\delta := \epsilon h$ in (23.4) with ϵ fixed small enough. Prove that $\inf_{f_h \in P_k(\mathcal{T}_h)} \|f - f_h\|_{L^p(D;\mathbb{R}^q)} \le c h^r \ell_D^{-r} \|f\|_{W^{r,p}(D;\mathbb{R}^q)}$ for all $r \in [0, k+1]$, all $p \in [1, \infty)$, and all $f \in W^{r,p}(D;\mathbb{R}^q)$. (*Hint*: admit as a fact that there is c, uniform, s.t. $\delta^s |\mathcal{K}_\delta(f)|_{W^{s,p}(D;\mathbb{R}^q)} \le c(\delta/\ell_D)^t \|f\|_{W^{t,p}(D;\mathbb{R}^q)}$ for all $s \ge t \ge 0$, then use $\mathcal{I}_h \circ \mathcal{K}_\delta$.)

Exercise 23.9 ($\boldsymbol{Z}_0^{\mathrm{c},p}(D) = \ker(\gamma^{\mathrm{c}})$). Let $p \in (1, \infty)$ and let $\boldsymbol{Z}_0^{\mathrm{c},p}(D) := \overline{\boldsymbol{\mathcal{C}}_0^\infty(D)}^{\boldsymbol{Z}^{\mathrm{c},p}(D)}$. We want to prove that $\boldsymbol{Z}_0^{\mathrm{c},p}(D) = \ker(\gamma^{\mathrm{c}})$ with the trace map $\gamma^{\mathrm{c}} : \boldsymbol{Z}^{\mathrm{c},p}(D) \to \boldsymbol{W}^{-\frac{1}{p},p}(\partial D)$ s.t.

$\langle \gamma^{\mathrm{c}}(\boldsymbol{v}), \boldsymbol{l}\rangle := \int_D \boldsymbol{v}\cdot\nabla\times\boldsymbol{w}(\boldsymbol{l})\,\mathrm{d}x - \int_D(\nabla\times\boldsymbol{v})\cdot\boldsymbol{w}(\boldsymbol{l})\,\mathrm{d}x$ for all $\boldsymbol{v}\in\boldsymbol{Z}^{\mathrm{c},p}(D)$ and all $\boldsymbol{l}\in\boldsymbol{W}^{\frac{1}{p},p'}(\partial D)$, where $\boldsymbol{w}(\boldsymbol{l})\in\boldsymbol{W}^{1,p}(D)$ is such that $\gamma^{\mathrm{d}}(\boldsymbol{w}(\boldsymbol{l})) = \boldsymbol{l}$ (see §4.3). (i) Show that $\boldsymbol{Z}_0^{\mathrm{c},p}(D)\subset\ker(\gamma^{\mathrm{c}})$. (*Hint*: $\mathcal{K}_\delta^{\mathrm{g}}(\boldsymbol{w})\to\boldsymbol{w}$ in $\boldsymbol{W}^{1,p}(D)$ as $\delta\to 0$ for all $\boldsymbol{w}\in\boldsymbol{W}^{1,p}(D)$ and $\gamma^{\mathrm{g}}: W^{1,p}(D)\to W^{\frac{1}{p},p'}(\partial D)$ is surjective.) (ii) Let $\boldsymbol{v}\in\ker(\gamma^{\mathrm{c}})$. Show that $\nabla\times\widetilde{\boldsymbol{v}} = \widetilde{\nabla\times\boldsymbol{v}}\in\boldsymbol{L}^p(\mathbb{R}^d)$, where for every function $\boldsymbol{v}$ defined in D, $\widetilde{\boldsymbol{v}}$ denotes its zero-extension to $\mathbb{R}^d$. (iii) Show that $\ker(\gamma^{\mathrm{c}})\subset\boldsymbol{Z}_0^{\mathrm{c},p}(D)$. (*Hint*: use the mollification operator $\mathcal{K}_{\delta,0}^{\mathrm{c}}$ defined in (23.23).)

Solution to exercises

Exercise 23.1 (Star-shaped domain). The smoothness properties of $\boldsymbol{\varphi}$ are evident, whereas (23.2) means that $(1-\delta)D + \delta B(\mathbf{0},r)\subset D$ which is nothing but the assumption that D is star-shaped with respect to the ball $B(\mathbf{0},r)$.

Exercise 23.2 (Commuting). Upon setting $\boldsymbol{T}(\boldsymbol{x}) := \boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y}$ for a fixed $\boldsymbol{y}\in B(\mathbf{0},1)$, the identities in Lemma 23.3 are simple consequences of the chain rule (see Lemma 9.6):

$$\begin{aligned}\nabla(f\circ\boldsymbol{T})(\boldsymbol{x}) &= \mathbb{J}_\delta^{\mathsf{T}}(\boldsymbol{x})(\nabla f)(\boldsymbol{T}(\boldsymbol{x})),\\ \nabla\cdot(\det(\mathbb{J}_\delta(\boldsymbol{x}))\mathbb{J}_\delta^{-1}(\boldsymbol{g}\circ\boldsymbol{T}))(\boldsymbol{x}) &= \det(\mathbb{J}_\delta(\boldsymbol{x}))(\nabla\cdot\boldsymbol{g})(\boldsymbol{T}(\boldsymbol{x})),\\ \nabla\times(\mathbb{J}_\delta^{\mathsf{T}}(\boldsymbol{x})(\boldsymbol{g}\circ\boldsymbol{T}))(\boldsymbol{x}) &= \det(\mathbb{J}_\delta(\boldsymbol{x}))\mathbb{J}_K^{-1}(\boldsymbol{x})(\nabla\times\boldsymbol{g})(\boldsymbol{T}(\boldsymbol{x})).\end{aligned}$$

Exercise 23.3 (Translation). (1) Assume first that f is smooth. Let $\boldsymbol{x}\in D$ and $v(t) := f(\boldsymbol{\mu}_{\lambda,t}(\boldsymbol{x}))$ with $t\in[0,1]$. The chain rule implies that $v'(t) = Df(\boldsymbol{\mu}_{\lambda,t}(\boldsymbol{x}))(\boldsymbol{\psi}_\lambda(\boldsymbol{x})-\boldsymbol{x})$, thereby showing that

$$f(\boldsymbol{\psi}_\lambda(\boldsymbol{x})) - f(\boldsymbol{x}) = \int_0^1 v'(t)\,\mathrm{d}t = \int_0^1 Df(\boldsymbol{\mu}_{\lambda,t}(\boldsymbol{x}))(\boldsymbol{\psi}_\lambda(\boldsymbol{x})-\boldsymbol{x})\,\mathrm{d}t.$$

Assuming that $p<\infty$, we infer that

$$\begin{aligned}\|f\circ\boldsymbol{\psi}_\lambda - f\|_{L^p(D)}^p &\le \int_D \|\boldsymbol{\psi}_\lambda(\boldsymbol{x})-\boldsymbol{x}\|_{\ell^2}^p \int_0^1 \|\nabla f(\boldsymbol{\mu}_{\lambda,t}(\boldsymbol{x}))\|_{\ell^2}^p\,\mathrm{d}t\,\mathrm{d}x\\ &\le c'\,\lambda^p\int_0^1\int_D \|\nabla f(\boldsymbol{\mu}_{\lambda,t}(\boldsymbol{x}))\|_{\ell^2}^p\,\mathrm{d}t\,\mathrm{d}x.\end{aligned}$$

The assumptions on $\boldsymbol{\psi}_\lambda$ imply that the mapping $\boldsymbol{\mu}_{\lambda,t}$ is invertible and $\|D\boldsymbol{\mu}_{\lambda,t}^{-1}\|_{\ell^2}\le 2$, $|\det(D\boldsymbol{\mu}_{\lambda,t}^{-1})|\le 2^d$. This gives

$$\|f\circ\boldsymbol{\psi}_\lambda - f\|_{L^p(D)}^p \le c'\,\lambda^p\int_0^1\int_D \|\nabla f(\boldsymbol{z})\|_{\ell^2}^p|\det(D\boldsymbol{\mu}_{\lambda,t}^{-1})|\,\mathrm{d}z\,\mathrm{d}t,$$

which finally implies that there is c_0 s.t. $\|f\circ\boldsymbol{\psi}_\lambda - f\|_{L^p(D)}\le c_0\,\lambda\|\nabla f\|_{\boldsymbol{L}^p(D)}$. The case $p=\infty$ is treated similarly.
(2) If f is not smooth, we deduce from Remark 23.8 that there exists a sequence of smooth functions converging to f in $W^{1,p}(D)$, i.e., for all $\epsilon>0$, there is a smooth function f_ϵ such that $\|f-f_\epsilon\|_{L^p(D)}\le\epsilon$ and $\|\nabla f_\epsilon\|_{\boldsymbol{L}^p(D)}\le 2\|\nabla f\|_{\boldsymbol{L}^p(D)}$. We infer that

$$\begin{aligned}\|f\circ\boldsymbol{\psi}_\lambda - f\|_{L^p(D)} &\le \|(f-f_\epsilon)\circ\boldsymbol{\psi}_\lambda\|_{L^p(D)} + \|f_\epsilon\circ\boldsymbol{\psi}_\lambda - f_\epsilon\|_{L^p(D)} + \|f_\epsilon - f\|_{L^p(D)}\\ &\le c\epsilon + 2c_0\lambda\|\nabla f\|_{\boldsymbol{L}^p(D)} + \epsilon.\end{aligned}$$

The conclusion follows readily since ϵ is arbitrary.

Exercise 23.4 (Approximation). (i) Let $f \in W^{s,p}(D)$ with $s \in (0,1)$ and $p \in [1,\infty)$. We estimate $\mathcal{K}^{\mathrm{g}}_\delta(f) - f$ in $L^p(D)$ as follows:

$$\begin{aligned}\|\mathcal{K}^{\mathrm{g}}_\delta(f) - f\|^p_{L^p(D)} &= \int_D \left| \int_{B(\mathbf{0},1)} \rho(y)\big(f(\boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y}) - f(\boldsymbol{x})\big)\,\mathrm{d}y\right|^p \mathrm{d}x \\ &\le c \int_{B(\mathbf{0},1)} \int_D \frac{|f(\boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y}) - f(\boldsymbol{x})|^p}{\|\boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y} - \boldsymbol{x}\|^{sp+d}_{\ell^2}} \|\boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y} - \boldsymbol{x}\|^{sp+d}_{\ell^2}\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Let us make the change of variables

$$B(\mathbf{0},1) \ni \boldsymbol{y} \mapsto \boldsymbol{z} = \boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y} \in \boldsymbol{\varphi}_\delta(D) + \delta r B(\mathbf{0},1) \subset D.$$

Observe that the Jacobian of this transformation is bounded from above by δr and

$$\|\boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y} - \boldsymbol{x}\|_{\ell^2} \le \|\boldsymbol{\varphi}_\delta(\boldsymbol{x}) - \boldsymbol{x}\|_{\ell^2} + \delta r\|\boldsymbol{y}\|_{\ell^2} \le c\,\delta.$$

Hence, we have

$$\|\mathcal{K}^{\mathrm{g}}_\delta(f) - f\|^p_{L^p(D)} \le c\,\delta^{sp+d}\delta^{-d} \int_D \int_D \frac{|f(\boldsymbol{z}) - f(\boldsymbol{x})|^p}{\|\boldsymbol{z} - \boldsymbol{x}\|^{sp+d}_{\ell^2}}\,\mathrm{d}x\,\mathrm{d}z \le c\,\delta^{sp}|f|^p_{W^{s,p}(D)}.$$

(ii) Assume now $s = 1$ and $p \in [1,\infty)$. Let $f \in W^{1,p}(D)$. By proceeding as above, we infer that

$$\|\mathcal{K}^{\mathrm{g}}_\delta(f) - f\|^p_{L^p(D)} \le c \int_{B(\mathbf{0},1)} \int_D |f(\boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y}) - f(\boldsymbol{x})|^p\,\mathrm{d}x\,\mathrm{d}y.$$

Let us fix $y \in B(\mathbf{0},1)$ and define the mapping $\boldsymbol{\psi}_\delta : D \ni \boldsymbol{x} \mapsto \boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y} \in \boldsymbol{\varphi}_\delta(D) + \delta r B(\mathbf{0},1) \subset D$. We observe that

$$\begin{aligned}&\|\boldsymbol{\psi}_\delta(\boldsymbol{x}) - \boldsymbol{x}\|_{\ell^2} \le \|\boldsymbol{\varphi}_\delta(\boldsymbol{x}) - \boldsymbol{x}\|_{\ell^2} + \delta r\|\boldsymbol{y}\|_{\ell^2} \le c\,\delta, \\ &\|D\boldsymbol{\psi}_\delta(\boldsymbol{x}) - \mathbb{I}\|_{\ell^2} = \|D\boldsymbol{\varphi}_\delta(\boldsymbol{x}) - \mathbb{I}\|_{\ell^2} \le c\,\delta,\end{aligned}$$

and $\boldsymbol{x} + t(\boldsymbol{\psi}_\delta(\boldsymbol{x}) - \boldsymbol{x}) = \boldsymbol{x} + t(\boldsymbol{\varphi}_\delta(\boldsymbol{x}) + \delta r\boldsymbol{y} - \boldsymbol{x}) \in D$, i.e., $\boldsymbol{\psi}_\delta$ satisfies the assumptions of Exercise 23.3. We infer that $\int_D |f(\boldsymbol{\varphi}_\delta(\boldsymbol{x}) + (\delta r)\boldsymbol{y}) - f(\boldsymbol{x})|^p\,\mathrm{d}x \le c\,\delta^p\|\nabla f\|^p_{\boldsymbol{L}^p(D)}$. We conclude that $\|\mathcal{K}^{\mathrm{g}}_\delta(f) - f\|_{L^p(D)} \le c\,\delta\|\nabla f\|_{\boldsymbol{L}^p(D)}$. The case $s = 1$, $p = \infty$ is treated similarly.
(iii) The definition (23.4) implies that $\mathcal{K}^{\mathrm{x}}_\delta(f) = \mathbb{K}^{\mathrm{x}}\mathcal{K}^{\mathrm{g}}_\delta(f)$ for all $\mathrm{x} \in \{\mathrm{c},\mathrm{d},\mathrm{b}\}$. Hence, we have

$$\begin{aligned}\|\mathcal{K}^{\mathrm{x}}_\delta(f) - f\|_{L^p(D;\mathbb{R}^q)} &\le \|\mathbb{K}^{\mathrm{x}}\mathcal{K}^{\mathrm{g}}_\delta(f) - \mathbb{K}^{\mathrm{x}}(f)\|_{L^p(D;\mathbb{R}^q)} + \|\mathbb{K}^{\mathrm{x}}(f) - f\|_{L^p(D;\mathbb{R}^q)} \\ &\le \|\mathbb{K}^{\mathrm{x}}\|_{L^\infty(D;\mathbb{R}^{q\times q})}\|\mathcal{K}^{\mathrm{g}}_\delta(f) - f\|_{L^p(D;\mathbb{R}^q)} + \|\mathbb{K}^{\mathrm{x}} - \mathbb{I}\|_{L^\infty(D;\mathbb{R}^{q\times q})}\|f\|_{L^p(D;\mathbb{R}^q)} \\ &\le c\,\big(\delta^s|f|_{W^{s,p}(D;\mathbb{R}^q)} + \ell_D^{-1}\delta\|f\|_{L^p(D;\mathbb{R}^q)}\big) \\ &\le c\,\ell_D^{-s}\delta^s\Big(\tfrac{\delta^{1-s}}{\ell_D^{1-s}}\|f\|_{L^p(D;\mathbb{R}^q)} + \ell_D^s|f|_{W^{s,p}(D;\mathbb{R}^q)}\Big).\end{aligned}$$

Since $\delta \le \ell_D$, this implies that $\|\mathcal{K}^{\mathrm{x}}_\delta(f) - f\|_{L^p(D)} \le c\ell_D^{-s}\delta^s\|f\|_{W^{s,p}(D)}$.

Exercise 23.5 (Preserving constants). Let us assume that $f \in Z^{\mathrm{g},p}(D)$. Let us consider $\check{\mathcal{K}}^{\mathrm{g}}_\delta(f) = \mathcal{K}^{\mathrm{g}}_\delta(f - \underline{f} - \underline{\nabla f}\cdot(\boldsymbol{x} - \boldsymbol{x}_D)) + \underline{f} + \underline{\nabla f}\cdot(\boldsymbol{x} - \boldsymbol{x}_D)$ as suggested in the hint, where $\underline{f} = \frac{1}{|D|}\int_D f\,\mathrm{d}x$ and $\underline{\nabla f} = \frac{1}{|D|}\int_D \nabla f\,\mathrm{d}x$. It is clear that if $f = \underline{f}$, then $\check{\mathcal{K}}^{\mathrm{g}}_\delta(f) = \underline{f} = f$, i.e., $\check{\mathcal{K}}^{\mathrm{g}}_\delta$ preserves constant fields. Moreover, $\check{\mathcal{K}}^{\mathrm{g}}_\delta(f) \in C^\infty(D;\mathbb{R})$. Notice also in passing that

$$\begin{aligned}\|\check{\mathcal{K}}^{\mathrm{g}}_\delta(f) - f\|_{L(D)} &\le \|\mathcal{K}^{\mathrm{g}}_\delta(f - \underline{f} - \underline{\nabla f}\cdot(\boldsymbol{x} - \boldsymbol{x}_D)) - (f - \underline{f} - \underline{\nabla f}\cdot(\boldsymbol{x} - \boldsymbol{x}_D))\|_{L(D)} \\ &\le c\,\ell_D^{-1}\delta\|f - \underline{f} - \underline{\nabla f}\cdot(\boldsymbol{x} - \boldsymbol{x}_D)\|_{W^{1,p}(D)} \\ &\le c\,\ell_D^{-1}\delta(\|f - \underline{f}\|_{W^{1,p}(D)} + \ell_D\|\underline{\nabla f}\|_{\boldsymbol{L}^p(D)}).\end{aligned}$$

The Poincaré-Steklov inequality implies that $\|\check{\mathcal{K}}_\delta^{\mathrm{g}}(f)-f\|_{L^p(D)}\le c\delta|f|_{W^{1,p}(D)}$.
Let us now modify $\mathcal{K}_\delta^{\mathrm{c}}$. Using that $\nabla\circ\mathcal{K}_\delta^{\mathrm{g}}=\mathcal{K}_\delta^{\mathrm{c}}\circ\nabla$, we have

$$\nabla\check{\mathcal{K}}_\delta^{\mathrm{g}}(f)=\mathcal{K}_\delta^{\mathrm{c}}(\nabla f-\underline{\nabla f})+\underline{\nabla f}.$$

Using that $\nabla\times(\boldsymbol{A}\times\boldsymbol{x})=2\boldsymbol{A}$ for all $\boldsymbol{A}\in\mathbb{R}^3$, the above identity suggests to take the following alternative definition for $\mathcal{K}_\delta^{\mathrm{c}}$:

$$\check{\mathcal{K}}_\delta^{\mathrm{c}}(\boldsymbol{g}):=\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{g}-\underline{\boldsymbol{g}}-\tfrac12\underline{\nabla\times\boldsymbol{g}}\times(\boldsymbol{x}-\boldsymbol{x}_D))+\underline{\boldsymbol{g}}+\tfrac12\underline{\nabla\times\boldsymbol{g}}\times(\boldsymbol{x}-\boldsymbol{x}_D).$$

Using $\boldsymbol{g}=\nabla f$ in this definition, we obtain $\nabla\check{\mathcal{K}}_\delta^{\mathrm{g}}(f)=\check{\mathcal{K}}_\delta^{\mathrm{c}}(\nabla f)$, i.e., the expected commutation holds true for all $f\in Z^{\mathrm{g},p}(D)$. Notice that $\check{\mathcal{K}}_\delta^{\mathrm{c}}(\boldsymbol{g})\in C^\infty(D;\mathbb{R}^3)$ and $\check{\mathcal{K}}_\delta^{\mathrm{c}}$ preserves constant fields. Proceeding as for $\check{\mathcal{K}}_\delta^{\mathrm{g}}$, we note in passing that we can also prove that $\|\check{\mathcal{K}}_\delta^{\mathrm{c}}(\boldsymbol{g})-\boldsymbol{g}\|_{\boldsymbol{L}^p(D)}\le c\delta|\boldsymbol{g}|_{\boldsymbol{W}^{1,p}(D)}$.
Let us continue with $\nabla\times\check{\mathcal{K}}_\delta^{\mathrm{c}}(\boldsymbol{g})$ where $\boldsymbol{g}\in\boldsymbol{Z}^{\mathrm{c},p}(D)$. Using that $(\nabla\times)\circ\mathcal{K}_\delta^{\mathrm{c}}=\mathcal{K}_\delta^{\mathrm{d}}\circ(\nabla\times)$, we have

$$\nabla\times\check{\mathcal{K}}_\delta^{\mathrm{c}}(\boldsymbol{g})=\mathcal{K}_\delta^{\mathrm{d}}(\nabla\times\boldsymbol{g}-\underline{\nabla\times\boldsymbol{g}})+\underline{\nabla\times\boldsymbol{g}}.$$

Using that $\nabla\cdot(\boldsymbol{A}\boldsymbol{x})=d\boldsymbol{A}$ for all $\boldsymbol{A}\in\mathbb{R}^d$, the above identity suggests to take

$$\check{\mathcal{K}}_\delta^{\mathrm{d}}\boldsymbol{g}:=\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{g}-\underline{\boldsymbol{g}}-\tfrac1d\underline{\nabla\cdot\boldsymbol{g}}(\boldsymbol{x}-\boldsymbol{x}_D))+\underline{\boldsymbol{g}}+\tfrac1d\underline{\nabla\cdot\boldsymbol{g}}(\boldsymbol{x}-\boldsymbol{x}_D).$$

Using $\boldsymbol{g}=\nabla\times\boldsymbol{h}$ in this definition, we obtain $\nabla\times\check{\mathcal{K}}_\delta^{\mathrm{c}}(\boldsymbol{h})=\check{\mathcal{K}}_\delta^{\mathrm{d}}(\nabla\times\boldsymbol{h})$, i.e., the expected commutation holds true. Notice that $\check{\mathcal{K}}_\delta^{\mathrm{d}}(\boldsymbol{g})\in C^\infty(D;\mathbb{R}^3)$ and $\check{\mathcal{K}}_\delta^{\mathrm{d}}$ preserves constant fields. Proceeding as for $\check{\mathcal{K}}_\delta^{\mathrm{g}}$, we can also prove that $\|\check{\mathcal{K}}_\delta^{\mathrm{d}}(\boldsymbol{g})-\boldsymbol{g}\|_{\boldsymbol{L}^p(D)}\le c\delta|\boldsymbol{g}|_{\boldsymbol{W}^{1,p}(D)}$.
Let us continue with $\nabla\cdot\check{\mathcal{K}}_\delta^{\mathrm{d}}(\boldsymbol{g})$ where $\boldsymbol{g}\in\boldsymbol{Z}^{\mathrm{d},p}(D)$. Using that $(\nabla\cdot)\circ\mathcal{K}_\delta^{\mathrm{d}}=\mathcal{K}_\delta^{\mathrm{b}}\circ\nabla\cdot$, we have

$$\nabla\cdot\check{\mathcal{K}}_\delta^{\mathrm{d}}(\boldsymbol{g})=\mathcal{K}_\delta^{\mathrm{b}}(\nabla\cdot\boldsymbol{g}-\underline{\nabla\cdot\boldsymbol{g}})+\underline{\nabla\cdot\boldsymbol{g}}.$$

Since we have reached the end of the de Rham diagram, we can set

$$\check{\mathcal{K}}_\delta^{\mathrm{b}}(f):=\mathcal{K}_\delta^{\mathrm{b}}(f-\underline{f})+\underline{f}.$$

This gives $\nabla\cdot\check{\mathcal{K}}_\delta^{\mathrm{d}}(\boldsymbol{g})=\check{\mathcal{K}}_\delta^{\mathrm{b}}(\nabla\cdot\boldsymbol{g})$, i.e., the expected commutation holds true. Notice that $\check{\mathcal{K}}_\delta^{\mathrm{b}}(f)\in C^\infty(D;\mathbb{R})$ and $\check{\mathcal{K}}_\delta^{\mathrm{b}}$ preserves constant fields. Proceeding as for $\check{\mathcal{K}}_\delta^{\mathrm{g}}$, we can also prove that $\|\check{\mathcal{K}}_\delta^{\mathrm{b}}(f)-f\|_{L^p(D)}\le c\delta|f|_{W^{1,p}(D)}$.

Exercise 23.6 (Inverse inequality). Let $\boldsymbol{x}\in K$. Since the function ρ is bounded, we infer that

$$\|\mathcal{K}_\delta(f)(\boldsymbol{x})\|_{\ell^2}\le c\int_{B(\mathbf{0},1)}\|f(\boldsymbol{\varphi}_{\delta(\boldsymbol{x})}(\boldsymbol{x})+\delta(\boldsymbol{x})\boldsymbol{y})\|_{\ell^2}\,\mathrm{d}y.$$

The condition (23.15b) implies that

$$\begin{aligned}\|\mathcal{K}_\delta(f)(\boldsymbol{x})\|_{\ell^2}&\le c\,\|\delta^{-1}\|_{L^\infty(D_K)}^d\int_{D_K}\|f(\boldsymbol{z})\|_{\ell^2}\,\mathrm{d}z\\&\le c\,\epsilon_{\min}^{-d}h_K^{-d}|D_K|^{1-\frac1p}\|f\|_{L^p(D_K;\mathbb{R}^q)}.\end{aligned}$$

We conclude using the regularity of the mesh sequence.

Exercise 23.7 (Approximation with $\mathcal{J}_h^{\mathrm{c}}$). Owing to the commuting property from Theorem 23.12, we infer that

$$\|\nabla\times(\boldsymbol{g}-\mathcal{J}_h^{\mathrm{c}}(\boldsymbol{g}))\|_{\boldsymbol{L}^p(D)}\le\|\nabla\times\boldsymbol{g}-\mathcal{J}_h^{\mathrm{d}}(\nabla\times\boldsymbol{g})\|_{\boldsymbol{L}^p(D)}.$$

Using the best-approximation result from Theorem 23.12 for x = c and x = d, and invoking Corollary 22.9 yields the desired bound.

Exercise 23.8 (Best approximation in L^p). Let $f \in L^p(D;\mathbb{R}^q)$ and consider the smallest integer l such that $W^{l,p}(D;\mathbb{R}^q)) \hookrightarrow V^{\mathrm{x}}(D)$ and $l \geq k+1$. The triangle inequality leads to

$$\|f - \mathcal{I}_h^{\mathrm{x}}(\mathcal{K}_\delta^{\mathrm{x}}(f))\|_{L^p(D;\mathbb{R}^q)} \leq \|f - \mathcal{K}_\delta^{\mathrm{x}}(f)\|_{L^p(D;\mathbb{R}^q)} + \|\mathcal{K}_\delta^{\mathrm{x}}(f) - \mathcal{I}_h^{\mathrm{x}}(\mathcal{K}_\delta^{\mathrm{x}}(f))\|_{L^p(D;\mathbb{R}^q)}.$$

We can bound $\|\mathcal{K}_\delta^{\mathrm{x}}(f) - \mathcal{I}_h^{\mathrm{x}}(\mathcal{K}_\delta^{\mathrm{x}}(f))\|_{L^p(D;\mathbb{R}^q)}$ by using the inverse inequality provided in the hint and Corollary 19.8 for x = g, Corollary 19.9 for x = c or Corollary 19.10 for x = d:

$$\begin{aligned}
\|\mathcal{K}_\delta^{\mathrm{x}}(f) - \mathcal{I}_h^{\mathrm{x}}(\mathcal{K}_\delta^{\mathrm{x}}(f))\|_{L^p(D;\mathbb{R}^q)}^p &= \sum_{K\in\mathcal{T}_h} \|\mathcal{K}_\delta^{\mathrm{x}}(f) - \mathcal{I}_h^{\mathrm{x}}(\mathcal{K}_\delta^{\mathrm{x}}(f))\|_{L^p(K;\mathbb{R}^q)}^p \\
&\leq c \sum_{K\in\mathcal{T}_h} \sum_{m\in\{k+1:\,l\}} h_K^{mp} |\mathcal{K}_\delta^{\mathrm{x}}(f)|_{W^{m,p}(K;\mathbb{R}^q)}^p \\
&\leq c \sum_{m\in\{k+1:\,l\}} h^{mp} |\mathcal{K}_\delta^{\mathrm{x}}(f)|_{W^{m,p}(D;\mathbb{R}^q)}^p \\
&\leq c\ell_D^{-rp} \|f\|_{W^{r,p}(D;\mathbb{R}^q)}^p \delta^{rp} \sum_{m\in\{k+1:\,l\}} h^{mp}\delta^{-mp},
\end{aligned}$$

which gives $\|\mathcal{K}_\delta^{\mathrm{x}}(f) - \mathcal{I}_h^{\mathrm{x}}(\mathcal{K}_\delta^{\mathrm{x}}(f))\|_{L^p(D;\mathbb{R}^q)} \leq c\ell_D^{-r} h^r \|f\|_{W^{r,p}(D;\mathbb{R}^q)}$ since $\delta = \epsilon h$ and ϵ is fixed. In conclusion, we have

$$\begin{aligned}
\|f - \mathcal{I}_h^{\mathrm{x}}(\mathcal{K}_\delta^{\mathrm{x}}(f))\|_{L^p(D;\mathbb{R}^q)} &\leq \|f - \mathcal{K}_\delta^{\mathrm{x}}(f)\|_{L^p(D;\mathbb{R}^q)} + c\ell_D^{-r} h^r \|f\|_{W^{r,p}(D;\mathbb{R}^q)} \\
&\leq c'\ell_D^{-r}(\delta^r + h^r)\|f\|_{W^{r,p}(D;\mathbb{R}^q)},
\end{aligned}$$

which gives the desired result (since $\delta = \epsilon h$ and ϵ is fixed).

Exercise 23.9 ($\boldsymbol{Z}_0^{\mathrm{c},p}(D) = \ker(\gamma^{\mathrm{c}})$). (i) Let $\boldsymbol{v} \in \boldsymbol{Z}_0^{\mathrm{c},p}(D)$. By definition, there is a sequence of smooth functions $(\boldsymbol{v}_n)_{n\in\mathbb{N}}$ in $\boldsymbol{C}_0^\infty(D)$ converging to $\boldsymbol{v}$ in $\boldsymbol{Z}^{\mathrm{c},p}(D)$. Let $\boldsymbol{l}$ be any function in $\boldsymbol{W}^{\frac{1}{p},p'}(\partial D)$. Since γ^{g} is surjective, there is a function $\boldsymbol{w}(\boldsymbol{l}) \in \boldsymbol{W}^{1,p}(D)$ such that $\gamma^{\mathrm{g}}(\boldsymbol{w}(\boldsymbol{l})) = \boldsymbol{l}$. Let $\delta > 0$ and let us consider $\mathcal{K}_\delta^{\mathrm{g}}(\boldsymbol{w}(\boldsymbol{l}))$. Since v_n is compactly supported, we have

$$\begin{aligned}
0 &= \int_D \nabla\cdot(\mathcal{K}_\delta^{\mathrm{g}}(\boldsymbol{w}(\boldsymbol{l}))\times\boldsymbol{v}_n)\,\mathrm{d}x \\
&= \int_D \boldsymbol{v}_n\cdot\nabla\times(\mathcal{K}_\delta^{\mathrm{g}}(\boldsymbol{w}(\boldsymbol{l})))\,\mathrm{d}x - \int_D (\mathcal{K}_\delta^{\mathrm{g}}(\boldsymbol{w}(\boldsymbol{l})))\cdot\nabla\times\boldsymbol{v}_n\,\mathrm{d}x.
\end{aligned}$$

Both integrals on the right-hand side converge as $\delta \to 0$ and $n \to \infty$. We infer that

$$\langle\gamma^{\mathrm{c}}(\boldsymbol{v}), \boldsymbol{l}\rangle := \int_D \boldsymbol{v}\cdot\nabla\times\boldsymbol{w}(\boldsymbol{l})\,\mathrm{d}x - \int_D \boldsymbol{w}\cdot\nabla\times\boldsymbol{v}(\boldsymbol{l})\,\mathrm{d}x = 0,$$

for every function $\boldsymbol{l}$ in $\boldsymbol{W}^{\frac{1}{p},p'}(\partial D)$. In conclusion, $\boldsymbol{v} \in \ker(\gamma^{\mathrm{c}})$. Hence, $\boldsymbol{Z}_0^{\mathrm{c},p}(D) \subset \ker(\gamma^{\mathrm{c}})$.
(ii) Let $\boldsymbol{v} \in \ker(\gamma^{\mathrm{c}})$. Let us show that $\widetilde{\boldsymbol{v}}$ has a weak curl in $\boldsymbol{L}^p(\mathbb{R}^d)$. Let $\boldsymbol{\phi} \in \boldsymbol{C}_0^\infty(\mathbb{R}^d)$. By definition, we have

$$\langle\nabla\times\widetilde{\boldsymbol{v}}, \boldsymbol{\phi}\rangle = \int_{\mathbb{R}^d} \widetilde{\boldsymbol{v}}\cdot\nabla\times\boldsymbol{\phi}\,\mathrm{d}x = \int_D \boldsymbol{v}\cdot\nabla\times\boldsymbol{\phi}\,\mathrm{d}x.$$

Using that $\boldsymbol{v} \in \ker(\gamma^{\mathrm{c}})$, the above equality implies that

$$\langle\nabla\times\widetilde{\boldsymbol{v}}, \boldsymbol{\phi}\rangle = \int_D \boldsymbol{v}\cdot\nabla\times\boldsymbol{\phi}\,\mathrm{d}x = \int_D \boldsymbol{\phi}\cdot\nabla\times\boldsymbol{v}\,\mathrm{d}x = \int_{\mathbb{R}^d} \boldsymbol{\phi}\cdot\widetilde{\nabla\times\boldsymbol{v}}\,\mathrm{d}x.$$

This proves that $\widetilde{\boldsymbol{v}}$ has a weak curl in $\boldsymbol{L}^p(\mathbb{R}^d)$ such that $\nabla\times\widetilde{\boldsymbol{v}} = \widetilde{\nabla\times\boldsymbol{v}}$. We have thus shown that $\boldsymbol{v} \in \widetilde{\boldsymbol{Z}}^{\mathrm{c},p}(D) := \{\boldsymbol{w} \in \boldsymbol{L}^p(D) \mid \widetilde{\nabla\times\boldsymbol{v}} \in \boldsymbol{L}^p(\mathbb{R}^d)\}$.
(iii) We can now apply the hint. According to Theorem 23.18 and Corollary 23.19, the sequence $(\mathcal{K}^{\mathrm{c}}_{\delta,0}(\boldsymbol{v}))_{\delta>0}$ converges to $\boldsymbol{v}$ in $\boldsymbol{Z}^{\mathrm{c},p}(D)$ (recall that we have established in Step (ii) that $\boldsymbol{v} \in \widetilde{\boldsymbol{Z}}^{\mathrm{c},p}(D)$). This proves that $\boldsymbol{v} \in \boldsymbol{Z}_0^{\mathrm{c},p}(D)$ since $\mathcal{K}^{\mathrm{c}}_{\delta,0}(\boldsymbol{v}) \in \boldsymbol{C}_0^\infty(D)$ (see Lemma 23.16). This shows that $\ker(\gamma^{\mathrm{c}}) \subset \boldsymbol{Z}_0^{\mathrm{c},p}(D)$.

Chapter 24

Weak formulation of model problems

Exercises

Exercise 24.1 (Forms). Let $D := (0,1)$. Which of these maps are linear or bilinear forms on $L^2(D)\times L^2(D)$: $a_1(f,g) := \int_D (f+g+1)\,\mathrm{d}x$, $a_2(f,g) := \int_D x(f-g)\,\mathrm{d}x$, $a_3(f,g) := \int_D (1+x^2)fg\,\mathrm{d}x$, $a_4(f,g) := \int_D (f+g)^2\,\mathrm{d}x$?

Exercise 24.2 ((Non)-uniqueness). Consider the domain D in $\mathbb{R}^2$ whose definition in polar coordinates is $D := \{(r,\theta) \mid r \in (0,1),\, \theta \in (\frac{\pi}{\alpha}, 0)\}$ with $\alpha \in (-1, -\frac{1}{2})$. Let $\partial D_1 := \{(r,\theta) \mid r = 1,\, \theta \in (\frac{\pi}{\alpha}, 0)\}$ and $\partial D_2 := \partial D \backslash \partial D_1$. Consider the PDE $-\Delta u = 0$ in D with the Dirichlet conditions $u = \sin(\alpha\theta)$ on ∂D_1 and $u = 0$ on ∂D_2. (i) Let $\varphi_1 := r^\alpha \sin(\alpha\theta)$ and $\varphi_2 := r^{-\alpha}\sin(\alpha\theta)$. Prove that φ_1 and φ_2 solve the above problem. (*Hint*: in polar coordinates $\Delta\varphi = \frac{1}{r}\partial_r(r\partial_r\varphi) + \frac{1}{r^2}\partial_{\theta\theta}\varphi$.) (ii) Prove that φ_1 and φ_2 are in $L^2(D)$ if $\alpha \in (-1, -\frac{1}{2})$. (iii) Consider the problem of seeking $u \in H^1(D)$ s.t. $u = \sin(\alpha\theta)$ on ∂D_1, $u = 0$ on ∂D_2, and $\int_D \nabla u \cdot \nabla v = 0$ for all $v \in H^1_0(D)$. Prove that φ_2 solves this problem, but φ_1 does not. Comment.

Exercise 24.3 (Poisson in 1D). Let $D := (0,1)$ and $f(x) := \frac{1}{x(1-x)}$. Consider the PDE $-\partial_x((1+\sin(x)^2)\partial_x u) = f$ in D with the Dirichlet conditions $u(0) = u(1) = 0$. Write a weak formulation of this problem with both trial and test spaces equal to $H^1_0(D)$ and show that the linear form on the right-hand side is bounded on $H^1_0(D)$. (*Hint*: notice that $f(x) = \frac{1}{x} + \frac{1}{1-x}$.)

Exercise 24.4 (Weak formulations). Prove Propositions 24.2 and 24.3.

Exercise 24.5 (Darcy). (i) Derive another variation on (24.12) and (24.14) with the functional spaces $V = W := \boldsymbol{H}(\mathrm{div}; D)\times L^2(D)$. (*Hint*: use Theorem 4.15.) (ii) Derive yet another variation with the functional spaces $V := \boldsymbol{L}^2(D)\times L^2(D)$ and $W := \boldsymbol{H}(\mathrm{div}; D)\times H^1_0(D)$.

Exercise 24.6 (Variational formulation). Prove that u solves (24.7) if and only if u minimizes over $H^1_0(D)$ the energy functional

$$\mathfrak{E}(v) := \frac{1}{2}\int_D |\nabla v|^2\,\mathrm{d}x - \int_D fv\,\mathrm{d}x.$$

(*Hint*: show first that $\mathfrak{E}(v+tw) = \mathfrak{E}(v) + t\left\{\int_D \nabla v\cdot\nabla w\,\mathrm{d}x - \int_D fw\,\mathrm{d}x\right\} + \frac{1}{2}t^2\int_D |\nabla w|^2\,\mathrm{d}x$ for all $v, w \in H^1_0(D)$ and all $t \in \mathbb{R}$.)

Exercise 24.7 (Derivative of primitive). Prove (24.18). (*Hint*: use Theorem 1.38 and Lebesgue's dominated convergence theorem.)

Exercise 24.8 (Biharmonic problem). Let D be an open, bounded, set in $\mathbb{R}^d$ with smooth boundary. Derive a weak formulation for the biharmonic problem

$$\Delta(\Delta u) = f \ \text{ in } D, \qquad u = \partial_n u = 0 \ \text{ on } \partial D,$$

with $f \in L^2(D)$. (*Hint*: use Theorem 3.16.)

Solution to exercises

Exercise 24.1 (Forms). The map a_1 is neither linear nor bilinear since $a_1(0,0) = |D| \neq 0$. The map a_2 is linear (not bilinear). The map a_3 is bilinear (not linear). The map a_4 is neither linear ($a_4(1,0) = |D| \neq \frac{1}{2}a_4(2,0) = 4|D|$) nor bilinear ($a_4(1,0) \neq 0$).

Exercise 24.2 ((Non)-uniqueness). (i) Direct verification.
(ii) We have $\varphi_2 \in L^\infty(D)$ since $\alpha < 0$, whereas φ_1 is in $L^2(D)$ if $2\alpha + 1 > -1$, i.e., if $\alpha > -1$.
(iii) One verifies that $\varphi_2 \in H^1(D)$ provided $2(-\alpha - 1) + 1 > -1$, i.e., $\alpha < 0$, which is indeed satisfied. The same argument shows that $\varphi_1 \notin H^1(D)$. This shows that by going from $L^2(D)$ to the smaller space $H^1(D)$, the nonuniqueness of the solution observed in Step (ii) disappears.

Exercise 24.3 (Poisson in 1D). We take $\varphi \in C_0^\infty(D)$, test the equation with φ, and integrate by parts. This yields

$$\int_0^1 d(x)\partial_x u \partial_x \varphi \,\mathrm{d}x = \int_0^1 f(x)\varphi(x)\,\mathrm{d}x,$$

with $d(x) := 1 + \sin(x)^2$. Notice that $\int_0^1 f(x)\varphi(x)\,\mathrm{d}x$ is unambiguously defined since f is of class C^∞ and bounded on the support of φ. We now want to pass to the limit and want to make sense of this integral when the test function is in $H_0^1(D)$. We have

$$\begin{aligned}\int_0^1 f(x)\varphi(x)\,\mathrm{d}x &= \int_0^1 \left(\frac{1}{x} + \frac{1}{1-x}\right)\varphi(x)\,\mathrm{d}x \\ &= \int_0^1 \partial_x\left(\ln(x) - \ln(1-x)\right)\varphi(x)\,\mathrm{d}x \\ &= -\int_0^1 \left(\ln(x) - \ln(1-x)\right)\partial_x\varphi(x)\,\mathrm{d}x.\end{aligned}$$

Hence, $g(x) = \ln(x) - \ln(1-x)$ is the weak derivative of f. Since $g \in L^2(D)$, the integral $-\int_0^1 g(x)\partial_x\varphi(x)\,\mathrm{d}x$ makes sense for all $\varphi \in H_0^1(D)$. In conclusion, the weak formulation consists of seeking $u \in H_0^1(D)$ such that

$$\int_0^1 d(x)\partial_x u \partial_x v \,\mathrm{d}x = -\int_0^1 g\partial_x v\,\mathrm{d}x, \qquad \forall v \in H_0^1(D).$$

Exercise 24.4 (Weak formulations). (i) Consider Proposition 24.2. Taking the test function $(\boldsymbol{\tau}, 0)$ with $\boldsymbol{\tau}$ arbitrary in $\boldsymbol{C}_0^\infty(D)$ shows that $\boldsymbol{\sigma} + \nabla p = 0$ a.e. in D. Take next the test function $(0, q)$ with q arbitrary in $C_0^\infty(D)$ to infer that $\nabla\cdot\boldsymbol{\sigma} = f$ a.e. in D. The boundary condition is explicitly enforced in the space for p.

(ii) Consider now Proposition 24.3. The PDE $\boldsymbol{\sigma} + \nabla p = 0$ a.e. in D is recovered as before. Take next the test function $(0, q)$ with q arbitrary in $C_0^\infty(D)$ to infer that $\boldsymbol{\sigma}$ has a weak divergence in $L^2(D)$ and that $\nabla{\cdot}\boldsymbol{\sigma} = f$ a.e. in D. The boundary condition on p is explicitly enforced.

Exercise 24.5 (Darcy). (i) Taking the functional spaces $V = W := \boldsymbol{H}(\mathrm{div}; D){\times}L^2(D)$ leads to the following weak formulation:

$$\begin{cases} \text{Find } u := (\boldsymbol{\sigma}, p) \in V \text{ such that} \\ \int_D (\boldsymbol{\sigma}{\cdot}\boldsymbol{\tau} - p\nabla{\cdot}\boldsymbol{\tau} - q\nabla{\cdot}\boldsymbol{\sigma})\,\mathrm{d}x = -\int_D fq\,\mathrm{d}x, \quad \forall w := (\boldsymbol{\tau}, q) \in W. \end{cases}$$

(ii) Taking the trial space $V := \boldsymbol{L}^2(D){\times}L^2(D)$ and the test space $W := \boldsymbol{H}(\mathrm{div}; D){\times}H_0^1(D)$ leads to the following weak formulation:

$$\begin{cases} \text{Find } u := (\boldsymbol{\sigma}, p) \in V \text{ such that} \\ \int_D (\boldsymbol{\sigma}{\cdot}\boldsymbol{\tau} - p\nabla{\cdot}\boldsymbol{\tau} - \boldsymbol{\sigma}{\cdot}\nabla q)\,\mathrm{d}x = \int_D fq\,\mathrm{d}x, \quad \forall w := (\boldsymbol{\tau}, q) \in W. \end{cases}$$

Exercise 24.6 (Variational formulation). The expanded formula for $\mathfrak{E}(v + tw)$ is established by developing the various terms and reordering them as zeroth-, first- and second-order terms in w. Let u solve (24.7). Taking $v := u$ and $t := 1$ in the expanded formula leads to $\mathfrak{E}(u + w) \geq \mathfrak{E}(u)$ for all $w \in \boldsymbol{H}_0^1(D)$. This implies that u minimizes $\mathfrak{E}$ over $H_0^1(D)$. Conversely, assume that u minimizes $\mathfrak{E}$ over $H_0^1(D)$. Let $w \in H_0^1(D)$. The right-hand side of the expanded formula is a second-order polynomial in t that is minimal at $t = 0$. Hence, the derivative of this polynomial vanishes at $t = 0$, which amounts to $\int_D \nabla u{\cdot}\nabla w\,\mathrm{d}x - \int_D fw\,\mathrm{d}x = 0$. Since w is arbitrary in $H_0^1(D)$, u solves (24.7).

Exercise 24.7 (Derivative of primitive). We use a density argument. Since $C_0^\infty(D)$ is dense in $L^1(D)$, there is a sequence $(f_n)_{n\in\mathbb{N}}$ in $C_0^\infty(D)$ that converges to f in $L^1(D)$ and such that $\|f_n\|_{L^1(D)} \leq 2\|f\|_{L^1(D)}$. Let $\phi \in C_0^\infty(D)$. It is clear that

$$\left| \int_0^1 f_n\phi\,\mathrm{d}s - \int_0^1 f\phi\,\mathrm{d}s \right| \leq (\sup_{x\in D} |\phi(x)|) \int_0^1 |f_n - f|\,\mathrm{d}s \to 0.$$

Likewise $(\int_0^x f_n\,\mathrm{d}s)\phi'(x) \to (\int_0^x f\,\mathrm{d}s)\phi'(x)$ a.e. in D, and

$$\left| \phi'(x) \int_0^x f_n\,\mathrm{d}s \right| \leq 2|\phi'(x)| \int_0^1 |f|\,\mathrm{d}s.$$

Lebesgue's dominated convergence theorem implies that

$$\int_0^1 (\int_0^x f_n\,\mathrm{d}s)\phi'(x)\,\mathrm{d}x \to \int_0^1 (\int_0^x f\,\mathrm{d}s)\phi'(x)\,\mathrm{d}x.$$

Passing to the limit in the relation

$$\int_0^1 \left(\int_0^x f_n\,\mathrm{d}s \right) \phi'(x)\,\mathrm{d}x = -\int_0^1 f_n(x)\phi(x)\,\mathrm{d}x$$

yields (24.18).

Exercise 24.8 (Biharmonic problem). Theorem 3.16 shows that

$$V := H_0^2(D) = \{v \in H^2(D) \mid v = \partial_n v = 0 \text{ on } \partial D\}.$$

Multiplying by a test function $v \in C_0^\infty(D)$, integrating over D, and using twice Green's formula along with the boundary conditions leads to the following weak formulation:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ \int_D \Delta u \Delta w \, \mathrm{d}x = \int_D f w \, \mathrm{d}x, \quad \forall w \in V. \end{cases}$$

If u solves the above weak formulation, taking $w \in C_0^\infty(D)$ shows that Δu has a weak Laplacian in $L^2(D)$, and since $f \in L^2(D)$, we infer that $\Delta(\Delta u) = f$ a.e. in D. The boundary conditions on u are explicitly enforced in V.

Chapter 25

Main results on well-posedness

Exercises

Exercise 25.1 (Riesz–Fréchet). The objective is to prove the Riesz–Fréchet theorem (Theorem C.24) by using the BNB theorem. Let V be a Hilbert space with inner product $(\cdot,\cdot)_V$. (i) Show that for every $v \in V$, there is a unique $J_V^{\mathrm{RF}}(v) \in V'$ s.t. $\langle J_V^{\mathrm{RF}}(v), w\rangle_{V',V} := (v,w)_V$ for all $w \in V$. (ii) Show that $J_V^{\mathrm{RF}} : V' \to V$ is a linear isometry.

Exercise 25.2 (Reflexivity). Let V, W be two Banach spaces such that there is an isomorphism $A \in \mathcal{L}(V;W)$. Assume that V is reflexive. Prove that W is reflexive. (*Hint*: consider the map $A^{**} \circ J_V \circ A^{-1}$.)

Exercise 25.3 (Space $V_{\mathbb{R}}$). Let V be a set and assume that V has a vector space structure over the field $\mathbb{C}$. By restricting the scaling λv to $\lambda \in \mathbb{R}$ and $v \in V$, V has also a vector space structure over the field $\mathbb{R}$, which we denote by $V_{\mathbb{R}}$ (V and $V_{\mathbb{R}}$ are the same sets, but they are equipped with different vector space structures); see Remark C.11. Let V' be the set of the bounded anti-linear forms on V and $V'_{\mathbb{R}}$ be the set of the bounded linear forms on $V_{\mathbb{R}}$. Prove that the map $I : V' \to V'_{\mathbb{R}}$ such that for all $\ell \in V'$, $I(\ell)(v) := \Re(\ell(v))$ for all $v \in V$, is a bijective isometry. (*Hint*: for $\psi \in V'_{\mathbb{R}}$, set $\ell(v) := \psi(v) + \mathrm{i}\psi(\mathrm{i}v)$ with $\mathrm{i}^2 = -1$.)

Exercise 25.4 (Orthogonal projection). Let V be a Hilbert space with inner product $(\cdot,\cdot)_V$ and induced norm $\|\cdot\|_V$. Let U be a nonempty, closed, and convex subset of V. Let $f \in V$. (i) Show that there is a unique u in U such that $\|f-u\|_V = \min_{v\in U}\|f-v\|_V$. (*Hint*: recall that $\frac{1}{4}(a-b)^2 = \frac{1}{2}(c-a)^2 + \frac{1}{2}(c-b)^2 - (c-\frac{1}{2}(a+b))^2$ and show that a minimizing sequence is a Cauchy sequence.) (ii) Show that $u \in U$ is the minimizer if and only if $\Re((f-u, v-u)_V) \le 0$ for all $v \in U$. (*Hint*: proceed as in the proof of Proposition 25.8.) (iii) Assuming that U is a (nontrivial) subspace of V, prove that the unique minimizer is characterized by $(f-u,v)_V = 0$ for all $v \in U$, and prove that the map $\Pi_U : V \ni f \mapsto u \in U$ is linear and $\|\Pi_U\|_{\mathcal{L}(V;U)} = 1$. (iv) Let a be a bounded, Hermitian, and coercive sesquilinear form (with $\xi := 1$ for simplicity). Let $\ell \in V'$. Set $\mathfrak{E}(v) := \frac{1}{2}a(v,v) - \ell(v)$. Show that there is a unique $u \in V$ such that $\mathfrak{E}(u) = \min_{v\in U}\mathfrak{E}(v)$ and that u is the minimizer if and only if $\Re(a(u, v-u) - \ell(v-u)) \ge 0$ for all $v \in U$.

Exercise 25.5 (Inf-sup constant). Let V be a Hilbert space, U a subset of V, and W a closed subspace of V. Let $\beta := \inf_{u\in U}\sup_{w\in W}\frac{|(u,w)_V|}{\|u\|_V\|w\|_W}$. (i) Prove that $\beta \in [0,1]$. (ii) Prove that

$\beta = \inf_{u\in U} \frac{\|\Pi_W(u)\|_V}{\|u\|_V}$, where Π_W is the orthogonal projection onto W. (*Hint*: use Exercise 25.4.) (iii) Prove that $\|u - \Pi_W(u)\|_V \le (1-\beta^2)^{\frac{1}{2}}\|u\|_V$. (*Hint*: use the Pythagorean identity.)

Exercise 25.6 (Fixed-point argument). The goal of this exercise is to derive another proof of the Lax–Milgram lemma. Let $A \in \mathcal{L}(V;V)$ be defined by $(A(v),w)_V := a(v,w)$ for all $v,w \in V$ (note that we use an inner product to define A). Let L be the representative in V of the linear form $\ell \in V'$. Let λ be a positive real number. Consider the map $T_\lambda : V \to V$ s.t. $T_\lambda(v) := v - \lambda\xi(A(v)-L)$ for all $v \in V$. Prove that if λ is small enough, $\|T_\lambda(v) - T_\lambda(w)\|_V \le \rho_\lambda \|v-w\|_V$ for all $v,w \in V$ with $\rho_\lambda \in (0,1)$, and show that (25.6) is well-posed. (*Hint*: use Banach's fixed-point theorem.)

Exercise 25.7 (Coercivity as necessary condition). Let V be a reflexive Banach space and let $A \in \mathcal{L}(V;V')$ be a monotone self-adjoint operator; see Definition C.31. Prove that A is bijective if and only if A is coercive (with $\xi := 1$). (*Hint*: prove that $\Re(\langle A(v),w\rangle_{V',V}) \le \langle A(v),v\rangle_{V',V}^{\frac{1}{2}} \langle A(w),w\rangle_{V',V}^{\frac{1}{2}}$ for all $v,w \in V$.)

Exercise 25.8 (Darcy). Prove that the problem (24.14) is well-posed. (*Hint*: adapt the proof of Proposition 25.18.)

Exercise 25.9 (First-order PDE). Prove that the problem (24.21) is well-posed. (*Hint*: adapt the proof of Proposition 25.19.)

Exercise 25.10 (T-coercivity). Let V, W be Hilbert spaces. Prove that (BNB1)-(BNB2) are equivalent to the existence of a bijective operator $T \in \mathcal{L}(V;W)$ and a real number $\eta > 0$ such that $\Re(a(v,T(v))) \ge \eta\|v\|_V^2$ for all $v \in V$. (*Hint*: use J_W^{-1}, $(A^{-1})^*$, and the map J_V^{RF} from the Riesz–Fréchet theorem to construct T.)

Exercise 25.11 (Sign-changing diffusion). Let D be a Lipschitz domain D in $\mathbb{R}^d$ partitioned into two disjoint Lipschitz subdomains D_1 and D_2. Set $\Sigma := \partial D_1 \cap \partial D_2$, each having an intersection with ∂D of positive measure. Let κ_1, κ_2 be two real numbers s.t. $\kappa_1 > 0$ and $\kappa_2 < 0$. Set $\kappa(x) := \kappa_1 \mathbb{1}_{D_1}(x) + \kappa_2 \mathbb{1}_{D_2}(x)$ for all $x \in D$. Let $V := H_0^1(D)$ be equipped with the norm $\|\nabla v\|_{\boldsymbol{L}^2(D)}$. The goal is to show that the bilinear form $a(v,w) := \int_D \kappa \nabla v \cdot \nabla w$ satisfies conditions (BNB1)-(BNB2) on $V \times V$; see Chesnel and Ciarlet [11]. Set $V_m := \{v|_{D_m} \mid v \in V\}$ for all $m \in \{1,2\}$, equipped with the norm $\|\nabla v_m\|_{\boldsymbol{L}^2(D_m)}$ for all $v_m \in V_m$, and let $\gamma_{0,m}$ be the traces of functions in V_m on Σ. (i) Assume that there is $S_1 \in \mathcal{L}(V_1;V_2)$ s.t. $\gamma_{0,2}(S_1(v_1)) = \gamma_{0,1}(v_1)$. Define $T : V \to V$ s.t. for all $v \in V$, $T(v)(x) := v(x)$ if $x \in D_1$ and $T(v)(x) := -v(x) + 2S_1(v|_{D_1})(x)$ if $x \in D_2$. Prove that $T \in \mathcal{L}(V)$ and that T is an isomorphism. (*Hint*: verify that $T \circ T = I_V$, the identity in V.) (ii) Assume that $\frac{\kappa_1}{|\kappa_2|} > \|S_1\|^2_{\mathcal{L}(V_1;V_2)}$. Prove that the conditions (BNB1)-(BNB2) are satisfied. (*Hint*: use T-coercivity from Remark 25.14.) (iii) Let $D_1 := (-a,0)\times(0,1)$ and $D_2 := (0,b)\times(0,1)$ with $a > b > 0$. Show that if $\frac{\kappa_1}{|\kappa_2|} \notin [1, \frac{a}{b}]$, (BNB1)-(BNB2) are satisfied. (*Hint*: consider the map $S_1 \in \mathcal{L}(V_1;V_2)$ s.t. $S_1(v_1)(x,y) := v_1(-\frac{a}{b}x, y)$ for all $v_1 \in V_1$, and the map $S_2 \in \mathcal{L}(V_2;V_1)$ s.t. $S_2(v_2)(x,y) := v_2(-x,y)$ if $x \in (-b,0)$ and $S_2(v_2)(x,y) := 0$ otherwise, for all $v_2 \in V_2$.)

Solution to exercises

Exercise 25.1 (Riesz–Fréchet). (i) For every $v \in V$, consider the linear form $\ell_v \in V'$ defined by $\ell_v(w) := (v,w)_V$ for all $w \in V$. Since $(\cdot,\cdot)_V$ is an inner product, we have

$$\|v\|_V = \sup_{w\in W} \frac{|(v,w)_V|}{\|w\|_V} = \sup_{w\in V} \frac{|\ell_v(w)|}{\|w\|_V}.$$

Hence, $\ell_v \in V'$ for all $v \in V$. Let us define the bilinear form $a(f,w) := \langle f, w\rangle_{V',V}$ for all $f \in V'$ and all $w \in V$. We need to prove that for all $v \in V$, there is a unique $J_V^{\mathrm{RF}}(v) \in V'$ s.t. $a(J_V^{\mathrm{RF}}(v), w) := \ell_v(w)$ for all $w \in V$. We do this by means of the BNB theorem, i.e., we verify the two assumptions of this theorem. We have

$$\|f\|_{V'} = \sup_{w\in V} \frac{|\langle f, w\rangle_{V',V}|}{\|w\|_V} = \sup_{w\in V} \frac{|a(f,w)|}{\|w\|_V}.$$

Hence, (25.11a) holds true:

$$\inf_{f\in V'} \sup_{w\in V} \frac{|a(f,w)|}{\|f\|_{V'}\|w\|_V} = 1.$$

Next, assume that $a(f,w) = 0$ for all $f \in V'$. Then $\langle f, w\rangle_{V',V} = 0$ for all $f \in V'$. Owing to Corollary C.14, we infer that

$$\|w\|_V = \sup_{f\in V'} \frac{|\langle f, w\rangle_{V',V}|}{\|f\|_{V'}} = 0.$$

This proves that $w = 0$. Hence, (25.11b) holds true as well.
(ii) $J_V^{\mathrm{RF}} : V' \to V$ is clearly linear. We have proved in Step (i) that

$$\|J_V^{\mathrm{RF}}(v)\|_{V'} = \sup_{w\in V} \frac{|a(f,w)|}{\|w\|_V} = \sup_{w\in V} \frac{|\ell_v(w)|}{\|w\|_V} = \|v\|_V.$$

Hence, $J_V^{\mathrm{RF}} : V' \to V$ is an isometry.

Exercise 25.2 (Reflexivity). Let us prove that $A^{**} \circ J_V \circ A^{-1} = J_W$. Let $w \in W$. We observe that

$$\begin{aligned}
\langle A^{**}(J_V(A^{-1}(w))), w'\rangle_{W'',W'} &= \langle J_V(A^{-1}(w), A^*(w'))\rangle_{V'',V'} \\
&= \overline{\langle A^*(w'), A^{-1}(w)\rangle_{V',V}} \\
&= \overline{\langle w', AA^{-1}(w)\rangle_{W',W}} \\
&= \overline{\langle w', w\rangle_{W',W}} = \langle J_W(w), w'\rangle_{W'',W'},
\end{aligned}$$

for all $w' \in W'$. Hence, $A^{**}(J_V(A^{-1}(w))) = J_W(w)$ in W', which proves that $A^{**} \circ J_V \circ A^{-1} = J_W$. Since A is an isomorphism, A^{**} is also an isomorphism (see Corollary C.52). Moreover, J_V is an isomorphism since V is reflexive. Hence, J_W is an isomorphism, which proves the reflexivity of W.

Exercise 25.3 (Space $V_{\mathbb{R}}$). The operator $I(\ell)$ maps onto $\mathbb{R}$ and is linear since $I(\ell)(tv) = \Re(\ell(tv)) = \Re(t\ell(v)) = t\Re(\ell(v)) = tI(\ell)(v)$ for all $t \in \mathbb{R}$ and all $v \in V$. Moreover, $I(\ell)$ is bounded since

$$|I(\ell)(v)| = |\Re(\ell(v))| \le |\ell(v)| \le \|\ell\|_{V'}\|v\|_V,$$

for all $v \in V$, so that $\|I(\ell)\|_{V'_{\mathbb{R}}} \le \|\ell\|_{V'}$. Furthermore, the map I is injective. Assuming indeed that $I(\ell) = 0$, i.e., $\Re(\ell(v)) = 0$ for all $v \in V$, the hypothesis $I(\ell) = 0$ implies that $0 = I(\ell)(iv) = \Re(\ell(\mathrm{i}v)) = \Re(-\mathrm{i}\ell(v)) = \Im(\ell(v))$ since $\mathrm{i}v \in V$. Hence, $I(\ell) = 0$ implies that $\Re(\ell(v)) = 0$ and $\Im(\ell(v)) = 0$. This proves that $\ell(v) = 0$ for all $v \in V$. Hence, $\ell = 0$. Let us now prove that I is surjective. Let $\psi \in V'_{\mathbb{R}}$ and consider the map $\ell : V \to \mathbb{C}$ s.t.

$$\ell(v) = \psi(v) + \mathrm{i}\psi(\mathrm{i}v), \qquad \forall v \in V.$$

(Recall that ψ is only $\mathbb{R}$-linear.) By construction, $I(\ell) = \psi$, and the map $\ell : V \to \mathbb{C}$ is antilinear. Indeed, let $\lambda \in \mathbb{C}$ and write $\lambda = \mu + \mathrm{i}\nu$ with $\mu, \nu \in \mathbb{R}$. We infer that

$$\begin{aligned}\ell(\lambda w) &= \psi(\mu w + \mathrm{i}\nu w) + \mathrm{i}\psi(\mathrm{i}\mu w - \nu w)\\ &= \mu\psi(w) + \nu\psi(\mathrm{i}w) + \mathrm{i}\mu\psi(\mathrm{i}w) - \mathrm{i}\nu\psi(w)\\ &= \mu\left(\psi(w) + \mathrm{i}\psi(iw)\right) - \mathrm{i}\nu\left(\psi(w) + \mathrm{i}\psi(\mathrm{i}w)\right) = \overline{\lambda}\ell(v),\end{aligned}$$

for all $w \in V$, where we used the $\mathbb{R}$-linearity of ψ. Let us finally show that $\|\ell\|_{V'} \le \|\psi\|_{V'_{\mathbb{R}}}$. Let $v \in V$ be s.t. $\ell(v) \neq 0$ and set $\lambda := \frac{\ell(v)}{|\ell(v)|} \in \mathbb{C}$. We have

$$|\ell(v)| = \lambda^{-1}\ell(v) = \ell(\lambda^{-1}v) = \psi(\lambda^{-1}v) + \mathrm{i}\psi(\mathrm{i}\lambda^{-1}v),$$

but since ψ takes values in $\mathbb{R}$, we infer that $\psi(\mathrm{i}\lambda^{-1}v) = 0$, so that $|\ell(v)| = \psi(\lambda^{-1}v)$. This implies that

$$|\ell(v)| \le \|\psi\|_{V'_{\mathbb{R}}}\|\lambda^{-1}v\|_V = \|\psi\|_{V'_{\mathbb{R}}}\|v\|_V,$$

since $|\lambda| = 1$. The proof is complete.

Exercise 25.4 (Orthogonal projection). (i) Let $(u_n)_{n\in\mathbb{N}}$ be a minimizing sequence in U. A direct calculation shows that

$$\begin{aligned}\frac{1}{4}\|u_n - u_m\|_V^2 &= \frac{1}{2}\|f - u_n\|_V^2 + \frac{1}{2}\|f - u_m\|^2 - \|f - \tfrac{1}{2}(u_n + u_m)\|_V^2\\ &\le \frac{1}{2}\|f - u_n\|_V^2 + \frac{1}{2}\|f - u_m\|^2,\end{aligned}$$

which shows that $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in V. Since V is a Hilbert space and U is closed, its limit, say u, is in U, and by construction, u is a minimizer. Uniqueness follows from the above formula since considering for u_n and u_m two minimizers gives

$$\|f - \tfrac{1}{2}(u_n + u_m)\|_V^2 < \frac{1}{2}\|f - u_n\|_V^2 + \frac{1}{2}\|f - u_m\|^2,$$

if u_n and u_m are distinct, which is a contradiction with the fact that u_n and u_m are minimizers.
(ii) We proceed as in the proof of Proposition 25.8 with $\mathfrak{E}(v) := \|f - v\|_V^2$. Let u be such that $\mathfrak{E}(u) = \min_{v\in U}\mathfrak{E}(v)$. Let $t \in [0,1]$. The formula (25.10) becomes

$$\mathfrak{E}(u + tw) = \mathfrak{E}(u) - 2t\Re((f - u, w)_V) + t^2\|w\|_V^2.$$

For all $v \in U$ and all $t \in [0,1]$, take $w := v - u$ and observe that $u + tw \in U$ by convexity of U. Hence, $\mathfrak{E}(u + tw) \ge \mathfrak{E}(u)$ for all $t \in [0,1]$. Since $p(t) := \mathfrak{E}(u + tw)$ is a polynomial in t and $p(t) \ge p(0)$ for all $t \in [0,1]$, we infer that its derivative at $t := 0$ is nonnegative, yielding $-\Re((f - u, v - u)_V) = p'(0) \ge 0$ for all $v \in U$. Conversely, if $u \in U$ satisfies this property, evaluating $\mathfrak{E}(u + tw)$ with $w := v - u$ at $t := 1$ yields $\mathfrak{E}(v) \ge \mathfrak{E}(u)$ for all $v \in U$.
(iii) When U is a subspace of V, $(v - u)$ spans U, so that the characterization becomes $\Re((f - u, v)_V) \le 0$ for all $v \in U$, and since the same inequality is verified for ξv for all $\xi \in \mathbb{C}$ with $|\xi| = 1$, we infer that $(f - u, v)_V = 0$ for all $v \in U$. The linearity of the map Π_U results from this characterization, which also yields

$$\|\Pi_U(f)\|_V^2 = \|f\|_V^2 - \|f - \Pi_U(f)\|_V^2,$$

for all $f \in V$, showing that $\|\Pi_U\|_{\mathcal{L}(V;U)} \le 1$. Equality is attained on the elements of U (unless $U = \{0\}$, in which case $\|\Pi_U\|_{\mathcal{L}(V;U)} = 0$).

(iv) All of the above can be extended by replacing the old functional $\|f-v\|_V^2$ by the new functional $\mathfrak{E}(v)$. In particular, we have

$$\frac{\alpha}{8}\|u_n-u_m\|_V^2 \le \frac{1}{8}a(u_n-u_m,u_n-u_m) = \frac{1}{2}\mathfrak{E}(u_n)+\frac{1}{2}\mathfrak{E}(u_m)-\mathfrak{E}(\tfrac{1}{2}(u_n+u_m)).$$

Exercise 25.5 (Inf-sup constant). (i) That $\beta \ge 0$ follows from its definition, and that $\beta \le 1$ follows from the Cauchy–Schwarz inequality.
(ii) The properties of the orthogonal projection imply that $\|\Pi_W(u)\|_V = \sup_{w\in W}\frac{|(u,w)_V|}{\|w\|_V}$. Let indeed r be the right-hand side. Taking $w := \Pi_W(u)$ in the supremum and since $(u,\Pi_W(u))_V = \|\Pi_W(u)\|_V^2$, we infer that $\|\Pi_W(u)\|_V \le r$. Moreover, since $(u,w)_V = (\Pi_W(u),w)_V$ for all $w\in W$, the Cauchy–Schwarz inequality implies that $|(u,w)_V| \le \|\Pi_W(u)\|_V\|w\|_V$. Hence, $r \le \|\Pi_W(u)\|_V$, and we conclude that equality holds true.
(iii) Step (ii) shows that $\|\Pi_W(u)\|_V \ge \beta\|u\|_V$ for all $u\in U$. Owing to the Pythagorean identity, we infer that

$$\|u-\Pi_W(u)\|_V^2 = \|u\|_V^2 - \|\Pi_W(u)\|_V^2 \le (1-\beta^2)\|u\|_V^2.$$

Exercise 25.6 (Fixed-point argument). We observe that

$$\begin{aligned}\|T_\lambda(v)-T_\lambda(w)\|_V^2 &= \|(v-w)-\lambda A(v-w)\|_V^2\\ &= \|v-w\|_V^2 - 2\lambda\Re(\xi a(v-w,v-w)) + \lambda^2\|A(v-w)\|_V^2\\ &\le (1-2\lambda\alpha+\lambda^2\|a\|_{V\times V}^2)\|v-w\|_V^2.\end{aligned}$$

Taking $\lambda := \frac{\alpha}{\|a\|_{V\times V}^2}$ yields $1-2\lambda\alpha+\lambda^2\|a\|_{V\times V}^2 = 1-(\frac{\alpha}{\|a\|_{V\times V}})^2 \in (0,1)$. The above bound shows that the map T_λ is a contraction. Owing to Banach's fixed-point theorem, there exists a unique $u\in V$ such that $T_\lambda(u)=u$, which is equivalent to u solving (25.6).

Exercise 25.7 (Coercivity as necessary condition). Assume that A is monotone and self-adjoint (so that $\langle Ay,y\rangle_{V',V}$ takes real values for all $y\in V$). Let $v,w\in V$. For all $t\in\mathbb{R}$, we infer that

$$\begin{aligned}0 &\le \langle A(v+tw),(v+tw)\rangle_{V',V}\\ &= \langle A(v),v\rangle_{V',V} + 2t\Re(\langle A(v),w\rangle_{V',V}) + t^2\langle A(w),w\rangle_{V',V}.\end{aligned}$$

The right-hand side is a second-order polynomial in t taking only nonnegative values. We infer that $\Re(\langle A(v),w\rangle_{V',V}) \le \langle A(v),v\rangle_{V',V}^{1/2}\langle A(w),w\rangle_{V',V}^{1/2}$. Assume now that A is bijective. The condition (BNB1) implies that there is a real number $\alpha>0$ such that for all $v\in V$,

$$\alpha\|v\|_V \le \sup_{w\in W}\frac{|\langle A(v),w\rangle_{W',W}|}{\|w\|_W} = \sup_{w\in W}\frac{\Re(\langle A(v),w\rangle_{W',W})}{\|w\|_W}.$$

Using the above bound on $\Re(\langle A(v),w\rangle_{V',V})$ yields

$$\alpha\|v\|_V \le \langle A(v),v\rangle_{V',V}^{\frac12}\sup_{w\in W}\frac{\langle A(w),w\rangle_{W',W}^{\frac12}}{\|w\|_W} \le \langle A(v),v\rangle_{V',V}^{\frac12}\|A\|_{\mathcal{L}(V;V')}^{\frac12}.$$

Hence, coercivity holds true with constant $\frac{\alpha^2}{\|A\|_{\mathcal{L}(V;V')}}$ and $\xi=1$.

Exercise 25.8 (Darcy). We verify conditions (BNB1)-(BNB2) for the bilinear form

$$a((\boldsymbol{\sigma},p),(\boldsymbol{\tau},q)) = \int_D (\boldsymbol{\sigma}\cdot\boldsymbol{\tau} + \nabla p\cdot\boldsymbol{\tau} + \boldsymbol{\sigma}\cdot\nabla q)\,\mathrm{d}x,$$

for the functional spaces $V = W := \boldsymbol{L}^2(D)\times H_0^1(D)$. Owing to the Poincaré–Steklov inequality for p, we equip these spaces with the norm $\|(\boldsymbol{\sigma},p)\|_V = \{\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}^2 + \|\nabla p\|_{\boldsymbol{L}^2(D)}^2\}^{1/2}$. Set

$$\mathbb{S} := \sup_{(\boldsymbol{\tau},q)\in V} \frac{a((\boldsymbol{\sigma},p),(\boldsymbol{\tau},q))}{\|(\boldsymbol{\tau},q)\|_V}.$$

We have

$$\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}^2 = a((\boldsymbol{\sigma},p),(\boldsymbol{\sigma},-p)) \le \mathbb{S}\,\|(\boldsymbol{\sigma},p)\|_V,$$
$$\|\nabla p\|_{\boldsymbol{L}^2(D)}^2 = a((\boldsymbol{\sigma},p),(\nabla p,-p)) \le 2\mathbb{S}\,\|(\boldsymbol{\sigma},p)\|_V,$$

so that (BNB1) holds true. Let now $(\boldsymbol{\tau},q)\in V$ be such that $a((\boldsymbol{\sigma},p),(\boldsymbol{\tau},q)) = 0$ for all $(\boldsymbol{\sigma},p)\in V$. Testing with $(\boldsymbol{\sigma},p) = (\boldsymbol{\tau},-q)$ yields $\|\boldsymbol{\tau}\|_{\boldsymbol{L}^2(D)}^2 = 0$, so that $\boldsymbol{\tau} = \boldsymbol{0}$. Moreover, testing with $(\boldsymbol{\sigma},p) := (\nabla q,-q)$ yields $\|\nabla q\|_{\boldsymbol{L}^2(D)}^2 = 0$, so that $q = 0$. Hence, (BNB2) holds true.

Exercise 25.9 (First-order PDE). Let $D := (0,1)$. We verify conditions (BNB1)-(BNB2) for the bilinear form

$$a(v,w) := \int_0^1 \frac{\mathrm{d}v}{\mathrm{d}t} w\,\mathrm{d}t,$$

and the functional spaces $V := \{v\in H^1(D)\mid v(0) = 0\}$ and $W := L^2(D)$ equipped with the norms $\|v\|_V := \|\frac{\mathrm{d}v}{\mathrm{d}t}\|_{L^2(D)}$ (this is legitimate owing to the Poincaré–Steklov inequality) and $\|w\|_W := \|w\|_{L^2(D)}$, respectively. We first observe that

$$\|v\|_V = \|\tfrac{\mathrm{d}v}{\mathrm{d}t}\|_{L^2(D)} = \sup_{w\in L^2(D)} \frac{|\int_D \frac{\mathrm{d}v}{\mathrm{d}t} w\,\mathrm{d}t|}{\|w\|_{L^2(D)}} = \sup_{w\in W} \frac{|a(v,w)|}{\|w\|_W},$$

so that (BNB1) holds true. Let now $w\in W$ be such that $\int_0^1 \frac{\mathrm{d}v}{\mathrm{d}t} w\,\mathrm{d}t = 0$ for all $v\in V$. Taking first $v\in C_0^\infty(D)$, we infer that w has a weak derivative such that $\frac{dw}{dt} = 0$. Hence, w is constant on D. Taking $v := t$ shows that $w = 0$, i.e., (BNB2) holds true.

Exercise 25.10 (T-coercivity). Assume the existence of a bijective operator $T\in\mathcal{L}(V;W)$ and a real number $\eta > 0$ such that $\Re(a(v,T(v))) \ge \eta\|v\|_V^2$ for all $v\in V$. The condition (BNB1) holds true with $\alpha := \frac{\eta}{\|T\|_{\mathcal{L}(V;W)}}$ since for all $v\in V$ with $v\neq 0$, we have

$$\eta\|v\|_V^2 \le \Re(a(v,T(v))) \le \frac{\Re(a(v,T(v)))}{\|T(v)|_W}\|T(v)\|_W$$
$$\le \sup_{w\in W}\frac{|a(v,w)|}{\|w\|_W}\|T\|_{\mathcal{L}(V;W)}\|v\|_V.$$

Let us show that the condition (BNB2) also holds true. Considering $w\in W$ such that $a(v,w) = 0$ for all $v\in V$, the bijectivity of T implies that there is $v_w\in V$ with $T(v_w) = w$. Taking v_w leads to $0 = a(v_w,w) = a(v_w,T(v_w))$, so that $0 = \Re(a(v_w,T(v_w))) \ge \eta\|v_w\|_V^2$. Hence, $v_w = 0$, so that $w = 0$.

Conversely, let us assume that (BNB1)-(BNB2) hold true. Then $A\in\mathcal{L}(V;W')$ is an isomorphism and $A^{-1}\in\mathcal{L}(W';V)$. Let us consider $(A^{-1})^*\in\mathcal{L}(V';W'')$. Set $T := J_W^{-1}\circ(A^{-1})^*\circ J_V^{\mathrm{RF}}$, where $J_W : W\to W''$ is the canonical isomorphism (recall that W is reflexive since it is a Hilbert space) and $J_V^{\mathrm{RF}} : V\to V'$ is the Riesz–Fréchet isomorphism. Then $T\in\mathcal{L}(V;W)$ and T is an isomorphism. Moreover, using that $\langle w', J_W^{-1}(w'')\rangle_{W',W} = \overline{\langle J_W(J_W^{-1}(w'')),w'\rangle_{W'',W'}} = \overline{\langle w'',w'\rangle_{W'',W'}}$ for all $w'\in W'$ and all $w''\in W''$, we infer that for all $v\in V$,

$$a(v,T(v)) = \langle A(v),T(v)\rangle_{W',W} = \overline{\langle (A^{-1})^*(J_V^{\mathrm{RF}}(v)),A(v)\rangle_{W'',W'}}$$
$$= \overline{\langle J_V^{\mathrm{RF}}(v),v\rangle_{V',V}} = \|v\|_V^2.$$

Exercise 25.11 (Sign-changing diffusion). (i) It is clear that the map T is linear and bounded. Moreover, we observe that $(T \circ T)(v)(x) = v(x)$ for all $v \in V$ and all $x \in D_1$, whereas we have for all $x \in D_2$,

$$\begin{aligned}(T \circ T)(v)(x) &= -T(v)(x) + 2S_1((T(v))|_{D_1})(x) \\ &= -(-v(x) + 2S_1(v|_{D_1})(x)) + 2S_1(v|_{D_1})(x) = v(x).\end{aligned}$$

Hence, $T \circ T = I_V$, and this implies that T is bijective.
(ii) To prove T-coercivity, we observe that for all $v \in V$,

$$\begin{aligned}a(v, T(v)) &= \int_{D_1} \kappa_1 |\nabla v|^2 \,\mathrm{d}x - \int_{D_2} \kappa_2 |\nabla v|^2 \,\mathrm{d}x + 2 \int_{D_2} \kappa_2 \nabla v \cdot \nabla (S_1(v|_{D_1})) \,\mathrm{d}x \\ &= \kappa_1 \|v|_{D_1}\|_{V_1}^2 - \kappa_2 \|v|_{D_2}\|_{V_2}^2 + 2 \int_{D_2} \kappa_2 \nabla v \cdot \nabla (S_1(v|_{D_1})) \,\mathrm{d}x \\ &\geq (\kappa_1 - \eta^{-1} |\kappa_2| \|S_1\|_{\mathcal{L}(V_1;V_2)}^2) \|v|_{D_1}\|_{V_1}^2 - \kappa_2 (1 - \eta) \|v|_{D_2}\|_{V_2}^2,\end{aligned}$$

where $\eta > 0$ can be chosen as small as needed. Under the assumption that $\frac{\kappa_1}{|\kappa_2|} > \|S_1\|_{\mathcal{L}(V_1;V_2)}^2$, it is possible to choose $\eta \in (0,1)$ such that both real numbers $(\kappa_1 - \eta^{-1}|\kappa_2| \|S_1\|_{\mathcal{L}(V_1;V_2)}^2)$ and $-\kappa_2(1-\eta)$ are positive. This yields T-coercivity, and, therefore, the conditions (BNB1)-(BNB2) are satisfied.
(iii) The map S_1 is in $\mathcal{L}(V_1; V_2)$, and one verifies that $\|S_1\|_{\mathcal{L}(V_1;V_2)}^2 = \frac{a}{b}$. Hence, the conditions (BNB1)-(BNB2) are satisfied if $\frac{\kappa_1}{|\kappa_2|} > \frac{a}{b}$. Reasoning similarly with the operator S_2 shows that T-coercivity holds true provided $\frac{\kappa_1}{|\kappa_2|} < \|S_2\|_{\mathcal{L}(V_2;V_1)}^2$, and one verifies that $\|S_2\|_{\mathcal{L}(V_2;V_1)}^2 = 1$.

Chapter 26

Basic error analysis

Exercises

Exercise 26.1 ((BNB2)). Prove that (26.8) is equivalent to (26.5b) provided (26.5a) holds true. (*Hint*: use that $\dim(W_h) = \operatorname{rank}(A_h) + \dim(\ker(A_h^*))$ (A_h^* is defined in (26.9)) together with the rank nullity theorem.)

Exercise 26.2 (Bijectivity of A_h^*). Prove that A_h is an isomorphism if and only if A_h^* is an isomorphism. (*Hint*: use $\dim(V_h) = \operatorname{rank}(A_h^*) + \dim(\ker(A_h))$ and $\dim(W_h) = \operatorname{rank}(A_h) + \dim(\ker(A_h^*))$.)

Exercise 26.3 (Petrov–Galerkin). Let V, W be real Hilbert spaces, let $A \in \mathcal{L}(V; W')$ be an isomorphism, and let $\ell \in W'$. Consider a conforming Petrov–Galerkin approximation with a finite-dimensional subspace $V_h \subset V$ and $W_h := (J_W^{\mathrm{RF}})^{-1} A V_h \subset W$, where $J_W^{\mathrm{RF}} : W \to W'$ is the Riesz–Fréchet isomorphism. The discrete bilinear form is $a_h(v_h, w_h) := \langle A(v_h), w_h \rangle_{W',W}$, and the discrete linear form is $\ell_h(w_h) := \ell(w_h)$ for all $v_h \in V_h$ and all $w_h \in W_h$. (i) Prove that the discrete problem (26.3) is well-posed. (ii) Show that its unique solution minimizes the residual functional $\mathfrak{R}(v) := \|A(v) - \ell\|_{W'}$ over V_h.

Exercise 26.4 (Fortin's lemma). (i) Prove that Π_h in the converse statement of Lemma 26.9 is idempotent. (*Hint*: prove that $B \circ A_h^{*\dagger} = I_{V_h'}$.) (ii) Assume that there are two maps $\Pi_{1,h}, \Pi_{2,h} : W \to W_h$ and two uniform constants $c_1, c_2 > 0$ such that $\|\Pi_{1,h}(w)\|_W \le c_1 \|w\|_W$, $\|\Pi_{2,h}((I - \Pi_{1,h})(w))\|_W \le c_2 \|w\|_W$ and $a(v_h, \Pi_{2,h}(w) - w) = 0$ for all $v_h \in V_h$, $w \in W$. Prove that $\Pi_h := \Pi_{1,h} + \Pi_{2,h}(I - \Pi_{1,h})$ is a Fortin operator. (iii) Write a variant of the direct statement in Lemma 26.9 assuming V, W reflexive, $A \in \mathcal{L}(V; W')$ bijective, and using this time an operator $\Pi_h : V \to V_h$ such that $a(\Pi_h(v) - v, w_h) = 0$ for all $(v, w_h) \in V \times W_h$ and $\gamma_{\Pi_h} \|\Pi_h(v)\|_V \le \|v\|_V$ for all $v \in V$ for some $\gamma_{\Pi_h} > 0$. (*Hint*: use (26.10) and Lemma C.53.)

Exercise 26.5 (Compact perturbation). Let V, W be Banach spaces with W reflexive. Let $A_0 \in \mathcal{L}(V; W')$ be bijective, let $T \in \mathcal{L}(V; W')$ be compact, and assume that $A := A_0 + T$ is injective. Let $a_0(v, w) := \langle A_0(v), w \rangle_{W',W}$ and $a(v, w) := \langle A(v), w \rangle_{W',W}$ for all $(v, w) \in V \times W$. Let $V_h \subset V$ and $W_h \subset W$ be s.t. $\dim(V_h) = \dim(W_h)$ for all $h \in \mathcal{H}$. Assume that approximability holds, and that the sesquilinear form a_0 satisfies the inf-sup condition

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a_0(v_h, w_h)|}{\|v_h\|_V \|w_h\|_W} =: \alpha_0 > 0, \qquad \forall h \in \mathcal{H}.$$

Following Wendland [46], the goal is to show that there is $h_0 > 0$ s.t.

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|w_h\|_W} =: \alpha > 0, \qquad \forall h \in \mathcal{H} \cap (0, h_0].$$

(i) Prove that $A \in \mathcal{L}(V; W')$ is bijective. (*Hint*: recall that a compact operator is bijective iff it is injective; this follows from the Fredholm alternative, Theorem 46.13.) (ii) Consider $R_h \in \mathcal{L}(V; V_h)$ s.t. for all $v \in V$, $R_h(v) \in V_h$ satisfies $a_0(R_h(v) - v, w_h) = 0$ for all $w_h \in W_h$. Prove that $R_h \in \mathcal{L}(V; V_h)$ and that $R_h(v)$ converges to v as $h \downarrow 0$ for all $v \in V$. (*Hint*: proceed as in the proof of Céa's lemma.) (iii) Set $L := I_V + A_0^{-1}T$ and $L_h := I_V + R_h A_0^{-1}T$ where I_V is the identity operator in V (observe that both L and L_h are in $\mathcal{L}(V)$). Prove that L_h converges to L in $\mathcal{L}(V)$. (*Hint*: use Remark C.5.) (iv) Show that if $h \in \mathcal{H}$ is small enough, L_h is bijective and there is C, independent of $h \in \mathcal{H}$, such that $\|L_h^{-1}\|_{\mathcal{L}(V)} \le C$. (*Hint*: observe that $L^{-1}L_h = I_V - L^{-1}(L - L_h)$ and consider the Neumann series.) (v) Conclude.

Solution to exercises

Exercise 26.1 ((BNB2)). The statement (26.8) is equivalent to $\ker(A_h^*) = \{0\}$. Since

$$\dim(W_h) = \operatorname{rank}(A_h) + \dim(\ker(A_h^*)),$$

this statement is equivalent to $\dim(W_h) = \operatorname{rank}(A_h)$. Since the inf-sup condition (26.5a) implies that $\ker(A_h) = \{0\}$, we infer that $\operatorname{rank}(A_h) = \dim(V_h)$ owing to the rank nullity theorem. We conclude that (26.8) is equivalent to $\dim(V_h) = \dim(W_h)$.

Exercise 26.2 (Bijectivity of A_h^*). We observe that $A_h : V_h \to W_h'$ is an isomorphism iff $\ker(A_h) = \{0\}$ and $\operatorname{rank}(A_h) = \dim(W_h') = \dim(W_h)$. Since we have $\dim(V_h) = \operatorname{rank}(A_h^*) + \dim(\ker(A_h))$ and $\dim(W_h) = \operatorname{rank}(A_h) + \dim(\ker(A_h^*))$, these two statements are equivalent to $\dim(V_h) = \operatorname{rank}(A_h^*)$ and $\dim(\ker(A_h^*)) = 0$, i.e., to the bijectivity of A_h^*.

Exercise 26.3 (Petrov–Galerkin). (i) We apply the discrete BNB theorem 26.6. Since $(J_W^{\mathrm{RF}})^{-1}A : V \to W$ is an isomorphism, the subspaces V_h and W_h have the same dimension, thus proving (26.5b). Moreover, taking $w_h := (J_W^{\mathrm{RF}})^{-1}A(v_h) \in W_h$ for all $v_h \in V_h$, we observe that

$$|a_h(v_h, w_h)| = |\langle A(v_h), (J_W^{\mathrm{RF}})^{-1}(A(v_h))\rangle_{W',W}| = \|A(v_h)\|_{W'}^2 \ge \alpha \|v_h\|_V^2,$$

for some real number $\alpha > 0$ since $A : V \to W'$ is an isomorphism. In addition, $\|w_h\|_W = \|A(v_h)\|_{W'} \le \|A\|_{\mathcal{L}(V;W')}\|v_h\|_V$. Combining these two bounds yields the inf-sup condition (26.5a). (ii) We observe that

$$\mathfrak{R}(v)^2 = \langle A(v) - \ell, (J_W^{\mathrm{RF}})^{-1}(A(v) - \ell)\rangle_{W',W}, \qquad \forall v \in V.$$

Proceeding as in the proof of Proposition 25.8, we infer that u_h minimizes $\mathfrak{R}$ over V_h iff

$$\langle A(u_h) - \ell, (J_W^{\mathrm{RF}})^{-1}A(v_h)\rangle_{W',W} = 0, \qquad \forall v_h \in V_h,$$

which is just a reformulation of the discrete problem in the Petrov–Galerkin setting.

Exercise 26.4 (Fortin's lemma). (i) We have for all $\theta_h \in V_h'$,

$$\begin{aligned}\langle B(A_h^{*\dagger}(\theta_h)), v_h\rangle_{V_h',V_h} &= \overline{a(v_h, A_h^{*\dagger}(\theta_h))} = \overline{\langle A_h(v_h), A_h^{*\dagger}(\theta_h)\rangle_{W_h',W_h}} \\ &= \langle A_h^*(A_h^{*\dagger}(\theta_h)), v_h\rangle_{V_h',V_h} = \langle \theta_h, v_h\rangle_{V_h',V_h},\end{aligned}$$

for all $v_h \in V_h$, which proves that $B \circ A_h^{*\dagger} = I_{V_h'}$. As a result, we have

$$\Pi_h(\Pi_h(w)) = A_h^{*\dagger}(B \circ A_h^{*\dagger}(B(w))) = A_h^{*\dagger}(B(w)) = \Pi_h(w),$$

for all $w \in W$, i.e., Π_h is idempotent.
(ii) We have

$$\begin{aligned} a(v_h, \Pi_h(w)) &= a(v_h, \Pi_{1,h}(w) + \Pi_{2,h}((I - \Pi_{1,h})(w))) \\ &= a(v_h, \Pi_{1,h}(w)) + a(v_h, \Pi_{2,h}((I - \Pi_{1,h})(w))) \\ &= a(v_h, \Pi_{1,h}(w)) + a(v_h, (I - \Pi_{1,h})(w)) = a(v_h, w). \end{aligned}$$

Moreover, we have

$$\|\Pi_h(w)\|_W \le \|\Pi_{1,h}(w)\|_W + \|\Pi_{2,h}((I - \Pi_{1,h})(w))\|_W \le (c_1 + c_2)\|w\|_W.$$

Hence, Π_h is a Fortin operator.
(iii) We observe that for all $w_h \in W_h$,

$$\sup_{v_h \in V_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V} \ge \sup_{v \in V} \frac{|a(\Pi_h(v), w_h)|}{\|\Pi_h(v)\|_V} \ge \gamma_{\Pi_h} \sup_{v \in V} \frac{|a(v, w_h)|}{\|v\|_V}.$$

Owing to Lemma C.53 (with W' in lieu of W) and since V is reflexive and A is an isomorphism, we infer that

$$\alpha = \inf_{v \in V} \sup_{w \in W} \frac{|a(v, w)|}{\|w\|_W \|v\|_V} = \inf_{w \in W} \sup_{v \in V} \frac{|a(v, w)|}{\|w\|_W \|v\|_V},$$

where we used the reflexivity of W and the fact that $a(v, w) = \langle A(v), w\rangle_{W',W}$. Since $w_h \in W_h \subset W$, this implies that $\sup_{v \in V} \frac{|a(v,w_h)|}{\|v\|_V} \ge \alpha \|w_h\|_W$, thereby proving that

$$\inf_{w_h \in W_h} \sup_{v_h \in V_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|w_h\|_W} \ge \gamma_{\Pi_h} \alpha.$$

Invoking (26.10), we conclude that a satisfies the inf-sup condition (26.5a) with $\alpha_h := \gamma_{\Pi_h}\alpha$.

Exercise 26.5 (Compact perturbation). (i) Since A_0 is bijective, we can write $A_0^{-1}A = I_V + A_0^{-1}T$. The operator $A_0^{-1}T$ being compact, the Fredholm alternative shows that $A_0^{-1}A$ is bijective if and only if it is injective. That $A_0^{-1}A$ is injective is a consequence of A_0 being bijective and A being injective.
(ii) Let $v \in V$. Owing to inf-sup condition satisfied by the sesquilinear form a_0, we infer that

$$\alpha_0 \|R_h(v)\|_V \le \sup_{w_h \in W_h} \frac{|a_0(R_h(v), w_h)|}{\|w_h\|_W} = \sup_{w_h \in W_h} \frac{|a_0(v, w_h)|}{\|w_h\|_W} \le \|a_0\| \, \|v\|_V,$$

where $\|a_0\|$ is the boundedness constant of a_0 on $V \times W$. Hence, $R_h \in \mathcal{L}(V; V_h)$ with $\|R_h\|_{\mathcal{L}(V;V_h)} \le \frac{\|a_0\|}{\alpha_0}$. Furthermore, proceeding as in the proof of Céa's lemma, we infer that

$$\|R_h(v) - v\|_V \le \left(1 + \frac{\|a_0\|}{\alpha_0}\right) \inf_{v_h \in V_h} \|v - v_h\|_V,$$

so that the convergence of $R_h(v)$ to v as $h \to 0$ follows from the approximability property.
(iii) Since R_h converges to I_V pointwise, we infer from Remark C.5 that R_h converges to I_V uniformly on compact sets. Since $A_0^{-1}T$ is compact, this implies that L_h converges to L uniformly

on bounded sets, i.e., in $\mathcal{L}(V)$.
(iv) Since $L^{-1}L_h = I_V - L^{-1}(L - L_h)$, taking $h \in \mathcal{H}$ small enough, say $h \in (0, h_1]$ with $h_0 > 0$, s.t. $\|L^{-1}(L - L_h)\|_{\mathcal{L}(V)} \leq \frac{1}{2}$, we infer that L_h is invertible with

$$L_h^{-1}L = \sum_{k\in\mathbb{N}} \big(L^{-1}(L - L_h)\big)^k,$$

so that $\|L_h^{-1}\|_{\mathcal{L}(V)} \leq C := 2\|L^{-1}\|_{\mathcal{L}(V)}$.
(v) Let $v_h \in V_h$. Assume $h \in (0, h_1]$. We infer that

$$\begin{aligned}\frac{\alpha_0}{C}\|v_h\|_V \leq \alpha_0\|L_h(v_h)\|_V &\leq \sup_{w_h\in W_h} \frac{|a_0(L_h(v_h), w_h)|}{\|w_h\|_W} \\ &\leq \sup_{w_h\in W_h} \frac{|a_0((L - L_h)(v_h), w_h)|}{\|w_h\|_W} + \sup_{w_h\in W_h} \frac{|a(v_h, w_h)|}{\|w_h\|_W},\end{aligned}$$

since $a_0(L(v_h), w_h) = \langle A_0(L(v_h)), w_h\rangle_{W',W} = \langle A(v_h), w_h\rangle_{W',W} = a(v_h, w_h)$. The first term on the right-hand side can be estimated by $\|a_0\|\,\|L - L_h\|_{\mathcal{L}(V)}\|v_h\|_V$, and taking $h \in \mathcal{H}$ small enough, say $h \in (0, h_2]$ with $h_2 > 0$, the factor $\|a_0\|\,\|L - L_h\|_{\mathcal{L}(V)}$ can be bounded by $\frac{\alpha_0}{2C}$ and thus hidden on the left-hand side, yielding the expected inf-sup condition for the sesquilinear form a with $h_0 := \min(h_1, h_2) > 0$.

Chapter 27

Error analysis with variational crimes

Exercises

Exercise 27.1 (Error identity). Assume stability, i.e., (27.1) holds true. Let $V_\sharp$ be defined in (27.2) and equip this space with a norm $\|\cdot\|_{V_\flat}$ s.t. there is $c_\flat$ s.t. $\|v_h\|_{V_\flat} \le c_\flat \|v_h\|_{V_h}$ for all $v_h \in V_h$. Prove that

$$\|u-u_h\|_{V_\flat} = \inf_{v_h\in V_h}\left[\|u-v_h\|_{V_\flat} + \frac{c_\flat}{\alpha_h}\|\delta_h(v_h)\|_{W_h'}\right].$$

Exercise 27.2 (Boundary penalty). (i) Prove that $x^2-2\beta xy+\eta_0 y^2 \ge \frac{\eta_0-\beta^2}{1+\eta_0}(x^2+y^2)$ for all real numbers x, y, $\eta_0\ge 0$ and $\beta\ge 0$. (ii) Using the notation of §27.3.1, prove that $a_h(v_h,v_h)\ge \frac38\|v_h\|_{V_h}^2$ for all $v_h\in V_h$. (*Hint*: prove that $|v_h'(0)v_h(0)| \le \|v_h'\|_{L^2(0,h)}h^{-\frac12}|v_h(0)|$.)

Exercise 27.3 (First-order PDE). The goal is to prove (27.11). (i) Prove that

$$h^{-\frac12}\|G(v_h)\|_{\ell^2(\mathbb{R}^I)} \le \sup_{w_h\in V_h}\frac{|a(v_h,w_h)|}{\|w_h\|_{L^2(D)}} \le \sqrt{6}h^{-\frac12}\|G(v_h)\|_{\ell^2(\mathbb{R}^I)},$$

where $G_i(v_h) := a(v_h,\varphi_i)$ for all $i\in\{1{:}I\}$ with $I:=\dim(V_h)$. (*Hint*: use Simpson's rule to compare Euclidean norms of component vectors and L^2-norms of functions.) (ii) Assume that I is even (the odd case is treated similarly). Prove that $\alpha_h\le c_2 h$. (*Hint*: consider the oscillating function v_h s.t. $v_h(x_{2i}):=2ih$ for all $i\in\{1{:}\frac{I}{2}\}$ and $v_h(x_{2i+1}):=1$ for all $i\in\{0{:}\frac{I}{2}-1\}$.) (iii) Prove that $\alpha_h\ge c_1 h$. (*Hint*: prove that $\max_{i\in\{1:I\}}|v_h(x_i)| \le 2\sum_{k\in\{1:I\}}|G_k(v_h)|$.) (iv) Prove that

$$\inf_{v_h\in V_h}\sup_{w_h\in W_h}\frac{|a(v_h,w_h)|}{\|v_h\|_{W^{1,1}(D)}\|w_h\|_{L^\infty(D)}} \ge \alpha_0 > 0$$

with $W_h := \{w_h\in L^\infty(D)\mid \forall i\in\{0{:}I-1\},\ w_{h|[x_i,x_{i+1}]}\in\mathbb{P}_0\}$. (*Hint*: see Proposition 25.19.)

Exercise 27.4 (GaLS 1D). The goal is to prove (27.12). Let $v_h\in V_h$. (i) Compute $a_h(v_h,v_h)$. (ii) Let $\zeta(x):=-2x/\ell_D$, set $\zeta_h:=\mathcal{I}_h^{\rm b}(\zeta)$, and show that $a_h(v_h,\mathcal{J}_h^{\rm av}(\zeta_h v_h)) \ge \frac12\ell_D^{-1}\|v_h\|_{L^2(D)}^2 -$

$c_1 a(v_h, v_h)$ uniformly w.r.t. $h \in \mathcal{H}$, $\mathcal{J}_h^{\mathrm{av}}$ is the averaging operator defined in (22.9), and $\mathcal{I}_h^{\mathrm{b}}$ is the L^2-projection on the functions that are piecewise constant over the mesh. (iii) Prove (27.12). (*Hint*: use the test function $z_h := 2\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h) + 2(c_1+1)v_h$.)

Exercise 27.5 (Nonconforming Strang 1). Let $T : W_h \to W \cap W_h$. Let $V_{\mathrm{S}} := V$ so that $V_\sharp := V + V_h$, and assume that $V_\sharp$ is equipped with a norm $\|\cdot\|_{V_\sharp}$ satisfying (27.5). (i) Assume that a_h can be extended to $V_h \times (W + W_h)$. Assume that there is $\|a\|_{\sharp h}$ s.t. consistency/boundedness holds true in the form $|a(u, T(w_h)) - a_h(v_h, T(w_h))| \le \|a\|_{\sharp h}\|u - v_h\|_{V_\sharp}\|w_h\|_{W_h}$. Prove that

$$\|u - u_h\|_{V_\sharp} \le \inf_{v_h \in V_h} \left[\left(1 + c_\sharp \frac{\|a\|_{\sharp h}}{\alpha_h}\right)\|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h}\|\hat{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'}\right],$$

with $\|\hat{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'} := \|\ell_h - \ell \circ T + a_h(v_h, T(\cdot)) - a_h(v_h, \cdot)\|_{W_h'}$. (*Hint*: add/subtract $a_h(v_h, T(w_h))$.) (ii) We now derive another error estimate that avoids extending a_h but restricts the discrete trial functions to $V_h \cap V$ (this is reasonable provided the subspace $V_h \cap V$ has approximation properties that are similar to those of V_h). Assuming that there is $\|a\|_{V \times W_h}$ s.t. boundedness holds true in the form $|a(u - v_h, T(w_h))| \le \|a\|_{V \times W_h}\|u - v_h\|_{V_\sharp}\|w_h\|_{W_h}$, prove that

$$\|u - u_h\|_{V_\sharp} \le \inf_{v_h \in V_h \cap V} \left[\left(1 + c_\sharp \frac{\|a\|_{V \times W_h}}{\alpha_h}\right)\|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h}\|\check{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'}\right],$$

with $\|\check{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'} := \|\ell_h - \ell \circ T + a(v_h, T(\cdot)) - a_h(v_h, \cdot)\|_{W_h'}$. (*Hint*: add/subtract $a(v_h, T(w_h))$.)

Exercise 27.6 (Orthogonal projection). Consider the setting of Exercise 25.4 with real vector spaces and coercivity with $\xi := 1$ for simplicity. Let u be the unique element in V such that $a(u, v - u) \ge \ell(v - u)$ for all $v \in U$. Let V_h be a finite-dimensional subspace of V, and let U_h be a nonempty, closed, and convex subset of V_h. We know from Exercise 25.4 that there is a unique u_h in V_h such that $a(u_h, v_h - u_h) \ge \ell(v_h - u_h)$ for all $v_h \in U_h$. (i) Show that there is $c_1(u)$ such that for all $(v, v_h) \in U \times V_h$,

$$\|u - u_h\|_V^2 \le c_1(u)\big(\|u - v_h\|_V + \|u_h - v\|_V + \|u - u_h\|_V\|u - v_h\|_V\big).$$

(*Hint*: prove $\alpha\|u - u_h\|_V^2 \le a(u, v - u_h) + \ell(u_h - v) + a(u_h, v_h - u) + \ell(u - v_h)$.) (ii) Show that there is $c_2(u)$ such that

$$\|u - u_h\|_V \le c_2(u)\Big(\inf_{v_h \in U_h}\big(\|u - v_h\|_V + \|u - v_h\|_V^2\big) + \inf_{v \in U}\|u_h - v\|_V\Big)^{\frac{1}{2}}.$$

Solution to exercises

Exercise 27.1 (Error identity). Let $v_h \in V_h$. The triangle inequality, the assumption on the $\|\cdot\|_{V_\flat}$-norm, stability, and the fact that the discrete solution satisfies $a_h(u_h, w_h) = \ell_h(w_h)$ for all

$w_h \in W_h$ imply that

$$\begin{aligned}\|u-u_h\|_{V_\flat} &\le \|u-v_h\|_{V_\flat} + \|u_h-v_h\|_{V_\flat}\\ &\le \|u-v_h\|_{V_\flat} + c_\flat\|u_h-v_h\|_{V_h}\\ &\le \|u-v_h\|_{V_\flat} + \frac{c_\flat}{\alpha_h}\sup_{w_h\in W_h}\frac{|a_h(u_h-v_h,w_h)|}{\|w_h\|_{W_h}}\\ &= \|u-v_h\|_{V_\flat} + \frac{c_\flat}{\alpha_h}\sup_{w_h\in W_h}\frac{|\ell_h(w_h)-a_h(v_h,w_h)|}{\|w_h\|_{W_h}}\\ &= \|u-v_h\|_{V_\flat} + \frac{c_\flat}{\alpha_h}\|\delta_h(v_h)\|_{W_h'}.\end{aligned}$$

Since v_h is arbitrary in V_h, we conclude that $\|u-u_h\|_{V_\flat} \le r_h$, where r_h denotes the right-hand side of the expected identity. Taking $v_h := u_h$ in the infimum and since $\delta_h(u_h) = 0 \in W_h'$, we conclude that $\|u-u_h\|_{V_\flat} \ge r_h$, i.e., $\|u-u_h\|_{V_\flat} = r_h$.

Exercise 27.2 (Boundary penalty). (i) Note that $x^2 - 2\beta xy + \eta_0 y^2 \ge \frac{\eta_0-\beta^2}{1+\eta_0}(x^2+y^2)$ iff

$$\frac{1+\beta^2}{1+\eta_0}x^2 - 2\beta xy + \frac{\eta^2+\beta^2}{1+\eta_0}y^2 \ge 0.$$

Since the coefficients of x^2 and y^2 are both positive, the above condition amounts to

$$\beta^2 \le \frac{1+\beta^2}{1+\eta_0}\frac{\eta^2+\beta^2}{1+\eta_0}.$$

Rearranging the terms leads to $2\eta_0\beta^2 \le \eta_0^2 + \beta^4$, which is trivially true.
(ii) Let $v_h \in V_h$. Since v_h' is piecewise constant, we have

$$|v_h'(0)v_h(0)| \le h^{\frac12}|v_h'(0)|\times h^{-\frac12}|v_h(0)| = \|v_h'\|_{L^2(0,h)}\times h^{-\frac12}|v_h(0)|.$$

Using the Cauchy–Schwarz inequality, we infer that

$$-v_h'(0)v_h(0) - v_h'(1)v_h(1) \ge -\|v_h'\|_{L^2(D)}\left(h^{-1}|v_h(0)|^2 + h^{-1}|v_h(1)|^2\right)^{\frac12}.$$

As a result, we have

$$a_h(v_h,v_h) \ge x^2 - xy + y^2,$$

with $x := \|v_h'\|_{L^2(D)}$ and $y := \left(h^{-1}|v_h(0)|^2 + h^{-1}|v_h(1)|^2\right)^{\frac12}$. Using the quadratic inequality from Step (i) with $\eta_0 := 1$ and $\beta := \frac12$, we conclude that

$$a_h(v_h,v_h) \ge \frac38(x^2+y^2) = \frac38\|v_h\|_{V_h}^2.$$

Exercise 27.3 (First-order PDE). (i) Denote by $\{\varphi_i\}_{i\in\{1:I\}}$ the nodal basis of V_h. Let $\mathsf{R}_\varphi : \mathbb{R}^I \to V_h$ be the isomorphism that reconstructs a function in V_h from its coordinate vector in $\mathbb{R}^I$. A direct computation using Simpson's rule (see §6.2) shows that $\frac16 h\|Y\|_{\ell^2(\mathbb{R}^I)}^2 \le \|\mathsf{R}_\varphi(Y)\|_{L^2(D)}^2 \le h\|Y\|_{\ell^2(\mathbb{R}^I)}^2$ for all $Y \in \mathbb{R}^I$. These bounds imply that

$$h^{-\frac12}\|G(v_h)\|_{\ell^2(\mathbb{R}^I)} \le \sup_{w_h\in V_h}\frac{|a(v_h,w_h)|}{\|w_h\|_{L^2(D)}} \le \sqrt{6}h^{-\frac12}\|G(v_h)\|_{\ell^2(\mathbb{R}^I)}, \tag{27.1}$$

where $G_i(v_h) := a(v_h, \varphi_i)$ for all $i \in \{1{:}I\}$.
(ii) Let us write $v_h := \sum_{i\in\{1:I\}} V_i\varphi_i$ with $V_{2i} := 2ih$ for all $i \in \{1{:}\frac{I}{2}\}$ and $V_{2i+1} := 1$ for all $i \in \{0{:}\frac{I}{2}{-}1\}$. Set $V_0 := 0$ and let $\lfloor\cdot\rfloor$ denote the floor function. We infer that

$$\begin{aligned} h\|v_h'\|_{L^2(D)}^2 &= \sum_{i\in\{0:I-1\}} h\int_{x_i}^{x_{i+1}} (v_h')^2 \,\mathrm{d}t = \sum_{i\in\{0:I-1\}} (V_{i+1}-V_i)^2 \\ &\geq \sum_{i\in\{0:\frac{I}{2}-1\}} (1-2ih)^2 \geq \sum_{i\in\{0:\lfloor\frac{I}{4}\rfloor\}} (1-2ih)^2 \geq \tfrac14(\lfloor\tfrac{I}{4}\rfloor+1), \end{aligned}$$

since $1-2ih \geq \frac12$ if $i \leq \lfloor\frac{I}{4}\rfloor$. Using the inequality $\lfloor\frac{I}{4}\rfloor + 1 > \frac{I}{4} = \frac{1}{4h}$ yields $\|v_h'\|_{L^2(D)} \geq \frac{1}{4h}$. Furthermore, $G_i(v_h) = 0$ if i is even, and $G_i(v_h) = h$ otherwise. Hence, $\|G(v_h)\|_{\ell^2(\mathbb{R}^I)} \leq h^{1/2}$. Using the rightmost bound in (27.1) leads to $\alpha_h \leq c_2 h$ with $c_2 := 4\sqrt{6}$.
(iii) Let v_h be arbitrary in V_h. Set $V_0 := 0$ and $V_{I+1} := V_I$. Since $G_i(v_h) = \frac12(V_{i+1} - V_{i-1})$ for all $i \in \{1{:}I\}$, we infer that for the even indices, we have $|V_{2i}| \leq 2\sum_{k\in\{0:i-1\}} |G_{2k+1}(v_h)|$, whereas for the odd indices, we have $|V_{2i-1}| \leq |V_{I+1}| + 2\sum_{k\in\{i:\frac{I}{2}\}} |G_{2k}(v_h)| \leq 2\sum_{k\in\{1:I\}} |G_k(v_h)|$. Hence, we have

$$\|V\|_{\ell^\infty(\mathbb{R}^I)} \leq 2\sum_{k\in\{1:I\}} |G_k(v_h)| \leq 2h^{-\frac12}\|G(v_h)\|_{\ell^2(\mathbb{R}^I)},$$

owing to the Cauchy–Schwarz inequality. Furthermore, a direct computation shows that $\|v_h'\|_{L^2(D)} \leq \sqrt{2}h^{-1}\|V\|_{\ell^\infty(\mathbb{R}^I)}$. Using the leftmost bound in (27.1) and the Poincaré inequality $\sqrt{2}\|v_h'\|_{L^2(D)} \geq \|v_h\|_{H^1(D)}$ leads to $\alpha_h \geq c_1 h$ with $c_2 := \frac14$.
(iv) Inspired by the proof of Proposition 25.19, we take $w_{h|[x_i,x_{i+1}]} := \mathrm{sgn}(v_{h|[x_i,x_{i+1}]}')$ for all $v_h \in V_h$ (notice that v_h' is piecewise constant). Then $\|w_h\|_{L^\infty(D)} = 1$ and $a(v_h, w_h) = \|v_h'\|_{L^1(D)}$, and we conclude as in Proposition 25.19.

Exercise 27.4 (GaLS 1D). Let $v_h \in V_h$.
(i) Applying the definition of $a_h(v_h, w_h)$, we obtain

$$a_h(v_h, v_h) = h\|v_h'\|_{L^2}^2 + \int_0^1 \frac12 (v_h^2)' \,\mathrm{d}x = h\|v_h'\|_{L^2(D)}^2 + \frac12 v_h(1)^2.$$

(ii) Let $\mathcal{J}_h^{\mathrm{av}}$ be the averaging operator defined in (22.9). Let $\zeta(x) = -2x/\ell_D$ and let us set $\zeta_h = \mathcal{I}_h^{\mathrm{b}}(\zeta)$. We have

$$\begin{aligned} a_h(v_h, \mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h)) &= a(v_h, \mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h)) + h\int_0^1 v_h'(\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h))' \,\mathrm{d}x \\ &= a(v_h, \zeta v_h) + a(v_h, (\zeta_h - \zeta)v_h) + a(v_h, \mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h) - \zeta_h v_h) \\ &\quad + h\int_0^1 v_h'(\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h))' \,\mathrm{d}x, \end{aligned}$$

where we recall that $a(v, w) := \int_0^1 v'w \,\mathrm{d}x$. Let us bound the four terms on the right-hand side. First, we have

$$a(v_h, \zeta v_h) = \int_0^1 v_h'\zeta v_h \,\mathrm{d}x = -\int_0^1 \frac12 v_h^2\zeta' \,\mathrm{d}x + \zeta(1)\frac12 v_h(1)^2 = \ell_D^{-1}\|v_h\|_{L^2(D)}^2 - v_h^2(1).$$

Second, using that $\|\zeta_h - \zeta\|_{L^\infty(D)} \leq c\ell_D^{-1}h$, we have

$$|a(v_h, (\zeta_h - \zeta)v_h)| \leq c\,\ell_D^{-1}h\|v_h'\|_{L^2(D)}\|v_h\|_{L^2(D)}.$$

Third, using (22.11) from Lemma 22.3, the fact that v_h and ζ are continuous and $\|\zeta_h-\zeta\|_{L^\infty(D)}\le c\,h\ell_D^{-1}$, and a discrete trace inequality, we have

$$\begin{aligned}|a(v_h,\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h)-\zeta_h v_h)| &\le \|v_h'\|_{L^2(D)}\|\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h)-\zeta_h v_h\|_{L^2(D)}\\ &\le c\,h^{\frac12}\|v_h'\|_{L^2(D)}\Big(\sum_{F\in\mathcal{F}_h^\circ}\|[\![\zeta_h v_h]\!]\|_{L^2(F)}^2\Big)^{\frac12}\\ &\le c\,h^{\frac12}\|v_h'\|_{L^2(D)}\Big(\sum_{F\in\mathcal{F}_h^\circ}\|[\![(\zeta_h-\zeta)]\!]v_h\|_{L^2(F)}^2\Big)^{\frac12}\\ &\le c\,\|\zeta_h-\zeta\|_{L^\infty(D)}\|v_h'\|_{L^2(D)}\|v_h\|_{L^2(D)}\\ &\le c\,\ell_D^{-1}\,h\|v_h'\|_{L^2}\|v_h\|_{L^2(D)}.\end{aligned}$$

Fourth, we use that ζ_h is piecewise constant, v_h is continuous, and we invoke (22.11) from Lemma 22.3 together with the triangle inequality, the bound $|\zeta_h v_h|_{H^1(K)}\le 2\ell_D^{-1}\|v_h\|_{L^2(K)}+2|v_h|_{H^1(K)}$, and the above manipulations on the jump term to infer that

$$\begin{aligned}h\int_0^1 v_h'(\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h))'\,\mathrm{d}x &\le h\|v_h'\|_{L^2(D)}\Big(\sum_{K\in\mathcal{T}_h}|\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h)|_{H^1(K)}^2\Big)^{\frac12}\\ &\le c\,h\|v_h'\|_{L^2(D)}\Big(\sum_{K\in\mathcal{T}_h}\ell_D^{-2}\|v_h\|_{L^2(K)}^2+\|v_h'\|_{L^2(K)}^2+h^{-1}\sum_{F\in\check{\mathcal{F}}_K^\circ}\|[\![\zeta_h-\zeta]\!]v_h\|_{L^2(F)}^2\Big)^{\frac12}\\ &\le c\,h\|v_h'\|_{L^2(D)}\Big(\sum_{K\in\mathcal{T}_h}\ell_D^{-2}\|v_h\|_{L^2(K)}^2+\|v_h'\|_{L^2(K)}^2+\ell_D^{-2}h\sum_{F\in\check{\mathcal{F}}_K^\circ}\|v_h\|_{L^2(F)}^2\Big)^{\frac12}\\ &\le c\,h\|v_h'\|_{L^2(D)}(\ell_D^{-1}\|v_h\|_{L^2}+\|v_h'\|_{L^2(D)}).\end{aligned}$$

In conclusion, we have established that

$$\begin{aligned}a_h(v_h,\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h)) &\ge \ell_D^{-1}\|v_h\|_{L^2(D)}^2-c\left(v_h(1)^2+\ell_D^{-1}h\|v_h'\|_{L^2(D)}\|v_h\|_{L^2(D)}+h\|v_h'\|_{L^2(D)}^2\right)\\ &\ge \frac12\ell_D^{-1}\|v_h\|_{L^2(D)}^2-c_1\left(\frac12 v_h(1)^2+h\|v_h'\|_{L^2(D)}^2\right),\end{aligned}$$

where the last bound follows from the use of Young's inequality.
(iii) The identity from Step (i) combined with the bound from Step (ii) implies that

$$a_h(v_h,\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h))+a_h(v_h,(c_1+1)v_h)\ge\frac12\Big(\ell_D^{-1}\|v_h\|_{L^2(D)}^2+v_h(1)^2+h\|v_h'\|_{L^2(D)}^2\Big).$$

Hence, $a_h(v_h,z_h)\ge\|v_h\|_{V_h}^2$ where $z_h:=2\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h)+2(c_1+1)v_h$. By proceeding as above, one can also show that $\|z_h\|_{V_h}\le c_2\|v_h\|_{V_h}$. This finally proves that

$$\sup_{w_h\in V_h}\frac{a_h(v_h,w_h)}{\|w_h\|_{V_h}}\ge\frac{a_h(v_h,z_h)}{\|z_h\|_{V_h}}\ge\frac{\|v_h\|_{V_h}^2}{\|z_h\|_{V_h}}\ge c_2^{-1}\|v_h\|_{V_h},$$

whence the inf-sup condition (27.12).

Exercise 27.5 (Nonconforming Strang 1). The starting point for both questions is the bound

$$\|u-u_h\|_{V_\sharp}\le\|u-v_h\|_{V_\sharp}+\frac{c_\sharp}{\alpha_h}\|\delta_h(v_h)\|_{W_h'}.$$

(i) Since a_h can be extended to $V_h \times (W + W_h)$, we can write

$$\begin{aligned}\langle \delta_h(v_h), w_h \rangle_{W_h', W_h} &= \ell_h(w_h) - a_h(v_h, w_h) \\ &= \ell_h(w_h) - \ell(T(w_h)) + a(u, T(w_h)) - a_h(v_h, w_h) \\ &\quad + a_h(v_h, T(w_h)) - a_h(v_h, T(w_h)) \\ &= \langle \hat{\delta}_h^{\mathrm{St1}}(v_h), w_h \rangle_{W_h', W_h} + a(u, T(w_h)) - a_h(v_h, T(w_h)),\end{aligned}$$

where $a(u, T(w_h)) = \ell(T(w_h))$ follows from $T(w_h) \in W$. Assuming consistency/boundedness in the form

$$|a(u, T(w_h)) - a_h(v_h, T(w_h))| \le \|a\|_{\sharp h} \, \|u - v_h\|_{V_\sharp} \|w_h\|_{W_h}$$

leads to

$$\|\delta_h(v_h)\|_{W_h'} \le \|\hat{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'} + \|a\|_{\sharp h} \, \|u - v_h\|_{V_\sharp},$$

whence we infer the expected error estimate.
(ii) Taking $v_h \in V_h \cap V$, we can write

$$\begin{aligned}\langle \delta_h(v_h), w_h \rangle_{W_h', W_h} &= \ell_h(w_h) - \ell(T(w_h)) + a(u, T(w_h)) - a_h(v_h, w_h) \\ &\quad + a(v_h, T(w_h)) - a(v_h, T(w_h)) \\ &= \langle \check{\delta}_h^{\mathrm{St1}}(v_h), w_h \rangle_{W_h', W_h} + a(u - v_h, T(w_h)).\end{aligned}$$

Assuming boundedness in the form

$$|a(u - v_h, T(w_h))| \le \|a\|_{V \times W_h} \, \|u - v_h\|_{V_\sharp} \|w_h\|_{W_h}$$

leads to

$$\|\delta_h(v_h)\|_{W_h'} \le \|\check{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'} + \|a\|_{V \times W_h} \, \|u - v_h\|_{V_\sharp},$$

whence we infer the expected error estimate.

Exercise 27.6 (Orthogonal projection). (i) Using the coercivity of a, we deduce that

$$\begin{aligned}\alpha \, \|u - u_h\|_V^2 &\le a(u - u_h, u - u_h) = a(u, u - u_h) - a(u_h, u - u_h) \\ &\le a(u, u - v) + a(u, v - u_h) - a(u_h, u - v_h) - a(u_h, v_h - u_h).\end{aligned}$$

Using that $a(u, u - v) \le \ell(u - v)$ and $a(u_h, u_h - v_h) \le \ell(u_h - v_h)$, the above inequality implies that

$$\begin{aligned}\alpha \|u - u_h\|_V^2 &\le \ell(u - v) + a(u, v - u_h) + a(u_h, v_h - u) + \ell(u_h - v_h) \\ &\le \ell(u - v_h) + a(u, v - u_h) + a(u_h, v_h - u) + \ell(u_h - v),\end{aligned}$$

which is the inequality suggested in the hint. Using the triangle inequality and letting $\|a\|$ denote the boundedness constant of a on $V \times V$, we infer that

$$\begin{aligned}\alpha \, \|u - u_h\|_V^2 &\le \|\ell\|_{V'} \|u - v_h\|_V + \|a\| \, \|u\|_V \|v - u_h\|_V \\ &\quad + \|a\| \, \|u_h\|_V \|v_h - u\|_V + \|\ell\|_{V'} \|u_h - v\| \\ &\le \|\ell\|_{V'} \|u - v_h\|_V + \|a\| \, \|u\|_V \|v - u_h\|_V \\ &\quad + \|a\| \, \|u_h - u\|_V \|v_h - u\|_V + \|a\| \, \|u\|_V \|v_h - u\|_V + \|\ell\|_{V'} \|u_h - v\| \\ &\le 2 \max(\|\ell\|_{V'}, \|a\| \, \|u\|_V)(\|u - v_h\|_V + \|v - u_h\|_V) + \|a\| \, \|u_h - u\|_V \|v_h - u\|_V.\end{aligned}$$

(ii) We split the quadratic term on the right-hand side as follows:

$$\|a\|\,\|u_h-u\|_V\|v_h-u\|_V \le \frac{\|a\|^2}{2\alpha}\|v_h-u\|_V^2+\frac{\alpha}{2}\|u-u_h\|_V^2.$$

We infer that

$$\frac{\alpha}{2}\|u-u_h\|_V^2 \le 2\max(\|\ell\|_V',\|a\|\,\|u\|_V)(\|u-v_h\|_V+\|v-u_h\|_V)+\frac{\|a\|^2}{2\alpha}\|v_h-u\|_V^2.$$

We can now take the infimum on $v_h\in U_h$ and on $v\in U$ since v_h and v are arbitrary.

Chapter 28

Linear algebra

Exercises

Exercise 28.1 (Matrix representation of operators). Let H be a (complex) Hilbert space with inner product $(\cdot,\cdot)_H$. Let V_h be a finite-dimensional subspace of H with basis $\{\varphi_i\}_{i\in\{1:I\}}$. Let $Z : V_h \to V_h$ be a linear operator. Let $\mathcal{M} \in \mathbb{C}^{I\times I}$ be the mass matrix s.t. $\mathcal{M}_{ij} := (\varphi_j, \varphi_i)_H$, and let $\mathcal{B}, \mathcal{D} \in \mathbb{C}^{I\times I}$ be s.t. $\mathcal{B}_{ij} := (Z(\varphi_j), \varphi_i)_H$, $\mathcal{D}_{ij} := (Z(\varphi_j), Z(\varphi_i))_H$ for all $i, j \in \{1:I\}$. Prove that $\mathcal{D} = \mathcal{B}^{\mathsf{H}}\mathcal{M}^{-1}\mathcal{B}$. (*Hint*: use $\mathcal{Z} \in \mathbb{C}^{I\times I}$ s.t. $Z(\varphi_j) := \sum_{k\in\{1:I\}} \mathcal{Z}_{kj}\varphi_k$.)

Exercise 28.2 (Smallest singular value). Prove that the real number α_{ℓ^2} defined (28.17a) is equal to $\|\mathcal{A}^{-1}\|^{-1}_{\ell^2(\mathbb{C}^I)}$. (*Hint*: to bound α_{ℓ^2}, consider a vector $\mathsf{V}_* \in \mathbb{C}^I$ s.t. $\|\mathcal{A}^{-1}\mathsf{V}_*\|_{\ell^2(\mathbb{C}^I)} = \|\mathcal{A}^{-1}\|_{\ell^2(\mathbb{C}^I)}\|\mathsf{V}_*\|_{\ell^2(\mathbb{C}^I)}$.)

Exercise 28.3 (ℓ^2-condition number). Let $\mathcal{Z} \in \mathbb{R}^{I\times I}$ be the upper triangular matrix such that $\mathcal{Z}_{ii} := 1$ for all $i \in \{1:I\}$, and $\mathcal{Z}_{ij} := -1$ for all $i, j \in \{1:I\}$, $i \neq j$. Let $\mathsf{X} \in \mathbb{R}^I$ have coordinates $\mathsf{X}_i := 2^{1-i}$ for all $i \in \{1:I\}$. Compute $\mathcal{Z}\mathsf{X}$, $\|\mathcal{Z}\mathsf{X}\|_{\ell^2(\mathbb{R}^I)}$, and $\|\mathsf{X}\|_{\ell^2(\mathbb{R}^I)}$. Show that $\|\mathcal{Z}\|_{\ell^2(\mathbb{R}^I)} \geq 1$ and derive a lower bound for $\kappa_{\ell^2}(\mathcal{Z})$. What happens if I is large?

Exercise 28.4 (Local mass matrix, 1D). Evaluate the local mass matrix for one-dimensional $\mathbb{P}_1$ and $\mathbb{P}_2$ Lagrange finite elements on a cell of length h.

Exercise 28.5 (Stiffness matrix). (i) Let $\{\widehat{\lambda}_1, \widehat{\lambda}_2, \widehat{\lambda}_3\}$ be the shape functions of the $\mathbb{P}_1$ Lagrange element with the cell $\widehat{K}$ shown on the leftmost part of Figure 28.1. Here, $\widehat{\lambda}_1$ is associated with the vertex $(1,0)$, $\widehat{\lambda}_2$ with the vertex $(0,1)$, and $\widehat{\lambda}_3$ with the vertex $(0,0)$. Evaluate the stiffness matrix for $\int_{\widehat{K}} \nabla v\cdot\nabla w \,\mathrm{d}x$. Same question for the $\mathbb{Q}_1$ Lagrange element with the shape functions $\{\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4\}$ associated with the vertices $(1,0), (1,1), (0,1), (0,0)$, respectively (see the central part of Figure 28.1). (ii) Consider the meshes of $D := (0,3)\times(0,2)$ shown in the right part of Figure 28.1. Evaluate the stiffness matrix for $\int_D \nabla v\cdot\nabla w \,\mathrm{d}x$.

Exercise 28.6 (Sensitivity to perturbations). Let $\mathcal{Z} \in \mathbb{C}^{I\times I}$ be invertible and let $\mathsf{X} \in \mathbb{C}^I$ solve $\mathcal{Z}\mathsf{X} = \mathsf{B}$ with $\mathsf{B} \neq 0$. Set $\kappa := \kappa_{\ell^2}(\mathcal{Z})$. (i) Let $\tilde{\mathsf{X}} \in \mathbb{C}^I$ solve $\mathcal{Z}\tilde{\mathsf{X}} = \tilde{\mathsf{B}}$. Prove that $\frac{\|\tilde{\mathsf{X}}-\mathsf{X}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathsf{X}\|_{\ell^2(\mathbb{C}^I)}} \leq \kappa\frac{\|\tilde{\mathsf{B}}-\mathsf{B}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathsf{B}\|_{\ell^2(\mathbb{C}^I)}}$. (ii) Let $\check{\mathsf{X}} \in \mathbb{C}^I$ solve $\check{\mathcal{Z}}\check{\mathsf{X}} = \mathsf{B}$. Prove that $\frac{\|\check{\mathsf{X}}-\mathsf{X}\|_{\ell^2(\mathbb{C}^I)}}{\|\check{\mathsf{X}}\|_{\ell^2(\mathbb{C}^I)}} \leq \kappa\frac{\|\check{\mathcal{Z}}-\mathcal{Z}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathcal{Z}\|_{\ell^2(\mathbb{C}^I)}}$. (iii) Explain why the above bounds are sharp.

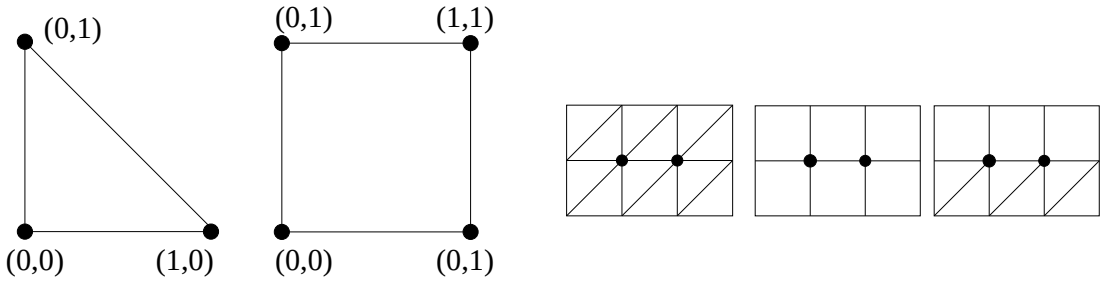


Figure 28.1: Illustration for Exercise 28.5. Left and central panels: reference triangle and square considered in Step (i). Right panel: three meshes for Step (ii).

Exercise 28.7 (Stability). Let $\mathcal{A}\mathsf{U} = \mathsf{B}$ be the linear system resulting from the Galerkin approximation. Equip the vector space $\mathbb{C}^I$ with the norm $\|\mathsf{V}\|_* := \sup_{\mathsf{Y}\in\mathbb{C}^I} \frac{|\mathsf{V}^\mathsf{H}\mathsf{Y}|}{\|\mathsf{R}_\psi(\mathsf{Y})\|_{W_h}}$. Show that $\frac{\|u_h - v_h\|_{V_h}}{\|u_h\|_{V_h}} \le \frac{\|a_h\|}{\alpha_h} \frac{\|\mathsf{B}-\mathcal{A}\mathsf{V}\|_*}{\|\mathsf{B}\|_*}$ for all $\mathsf{V} \in \mathbb{C}^I$, where $u_h := \mathsf{R}_\varphi(\mathsf{U})$ and $v_h := \mathsf{R}_\varphi(\mathsf{V})$. (*Hint*: show that $\alpha_h\|u_h - v_h\|_{V_h} \le \|\mathcal{A}(\mathsf{U}-\mathsf{V})\|_*$ and that $\|\mathsf{B}\|_* \le \|a_h\|\|u_h\|_{V_h}$, where α_h and $\|a_h\|$ are the stability and boundedness constants of a_h on $V_h \times W_h$.)

Exercise 28.8 (ℓ^∞-norm). (i) Prove Proposition 28.18. (*Hint*: use that $\mathcal{A}\mathsf{Y} \ge \min_{j\in\{1:I\}}(\mathcal{A}\mathsf{Y})_j\mathsf{U}$, where $\mathsf{U} \in \mathbb{R}^I$ has all entries equal to 1.) (ii) Derive a bound on $\|\mathcal{A}^{-1}\|_{\ell^\infty(\mathbb{R}^I)}$ with $\mathcal{A} := h^{-1}\,\mathrm{tridiag}(-1,2,-1)$. (*Hint*: consider the function $x \mapsto x(1-x)$ on $(0,1)$ to build a majorizing vector.) (iii) Let $(\mathsf{E}_1,\dots,\mathsf{E}_I)$ be the canonical basis of $\mathbb{R}^I$. Let $\alpha \in \mathbb{R}$ and consider the matrix $\mathcal{Z} := \mathcal{I} + \alpha\mathsf{E}_1 \otimes \mathsf{E}_I$ with entries $\mathcal{Z}_{ij} := \delta_{ij} + \alpha\delta_{i1}\delta_{jI}$. Verify that $\mathcal{Z}^{-1} = \mathcal{I} - \alpha\mathsf{E}_1 \otimes \mathsf{E}_I$ and evaluate the condition number $\kappa_{\ell^\infty}(\mathcal{Z})$. What happens if α is large?

Exercise 28.9 (Lumped mass matrix). Let D be a two-dimensional polygonal set and consider an affine mesh $\mathcal{T}_h$ of D composed of triangles and $\mathbb{P}_1$ Lagrange elements. (i) Let K be a cell in $\mathcal{T}_h$. Compute the local mass matrix $\mathcal{M}^K$ with entries $\mathcal{M}^K_{ij} := \int_K \theta_{K,i}(\boldsymbol{x})\theta_{K,j}(\boldsymbol{x})\,\mathrm{d}x$, $i,j \in \{1{:}3\}$. (ii) Compute the lumped local mass matrix $\overline{\mathcal{M}}^K$ with $\overline{\mathcal{M}}^K_{ij} := \delta_{ij}\sum_{l\in\{1:3\}}\mathcal{M}^K_{il}$. (iii) Compute the eigenvalues of $(\overline{\mathcal{M}}^K)^{-1}(\overline{\mathcal{M}}^K - \mathcal{M}^K)$. (iv) Let $\mathcal{M}$ be the global mass matrix and $\overline{\mathcal{M}}$ be the lumped mass matrix. Show that the largest eigenvalue of $(\overline{\mathcal{M}})^{-1}(\overline{\mathcal{M}} - \mathcal{M})$ is $\frac{3}{4}$.

Exercise 28.10 (CG). Let $\mathcal{A} \in \mathbb{R}^{I\times I}$ be a real symmetric positive definite matrix and let $\mathfrak{J} : \mathbb{R}^I \to \mathbb{R}$ be such that $\mathfrak{J}(\mathsf{V}) := \frac{1}{2}\mathsf{V}^\mathsf{T}\mathcal{A}\mathsf{V} - \mathsf{B}^\mathsf{T}\mathsf{V}$. Let U_m be the iterate at step $m \ge 1$ of the CG method. (i) Prove that U_m minimizes $\mathfrak{J}$ over $\mathsf{U}_0 + K_m$. (*Hint*: use Proposition 28.20.) (ii) Let $\eta_m := \arg\min_{\eta\in\mathbb{C}} \mathfrak{J}(\mathsf{U}_m + \eta\mathsf{P}_m)$. Show that $\eta_m = \alpha_m$ in the CG method. (iii) Write the preconditioned CG method by just invoking the matrix $\mathcal{P} := \mathcal{P}_\mathrm{L}\mathcal{P}_\mathrm{L}^\mathsf{T}$.

Exercise 28.11 (Complex symmetric system). Let $\mathcal{A} := \mathcal{T} + \mathrm{i}\sigma\mathcal{I}$ where $\mathcal{T}$ is symmetric real, $\sigma > 0$, and $\mathcal{I}$ is the identity matrix of size $I\times I$. Let $\mathcal{A}_*$ and $\mathcal{A}_{**}$ be the two rewritings of $\mathcal{A}$ as a real matrix of size $2I\times 2I$ (see Remark 28.23). Determine the spectra $\sigma(\mathcal{A})$, $\sigma(\mathcal{A}_*)$, and $\sigma(\mathcal{A}_{**})$, and comment on their position with respect to the origin. What happens if one considers the rotated linear system $-\mathrm{i}\mathcal{A}U = -\mathrm{i}B$ instead?

Solution to exercises

Exercise 28.1 (Matrix representation). Let $i, j \in \{1{:}I\}$. We have

$$\mathcal{B}_{ij} = (Z(\varphi_j), \varphi_i)_H = \sum_{k\in\{1:I\}} \mathcal{Z}_{kj}(\varphi_k, \varphi_i)_H = \sum_{k\in\{1:I\}} \mathcal{Z}_{kj}\mathcal{M}_{ik} = (\mathcal{M}\mathcal{Z})_{ij},$$

i.e., $\mathcal{B} = \mathcal{M}\mathcal{Z}$. Moreover, we have

$$\begin{aligned}\mathcal{D}_{ij} = (Z(\varphi_j), Z(\varphi_i))_H &= \sum_{k\in\{1:I\}}\sum_{l\in\{1:I\}} \mathcal{Z}_{kj}\overline{\mathcal{Z}_{li}}(\varphi_k, \varphi_l)_H \\ &= \sum_{k\in\{1:I\}}\sum_{l\in\{1:I\}} \mathcal{Z}_{kj}\mathcal{Z}_{il}^{\mathsf{H}}\mathcal{M}_{lk} = (\mathcal{Z}^{\mathsf{H}}\mathcal{M}\mathcal{Z})_{ij},\end{aligned}$$

showing that $\mathcal{D} = \mathcal{Z}^{\mathsf{H}}\mathcal{M}\mathcal{Z}$. Putting everything together, and since $\mathcal{M}^{\mathsf{H}} = \mathcal{M}$, we conclude that

$$\mathcal{D} = (\mathcal{M}^{-1}\mathcal{B})^{\mathsf{H}}\mathcal{M}(\mathcal{M}^{-1}\mathcal{B}) = \mathcal{B}^{\mathsf{H}}\mathcal{M}^{-1}\mathcal{B}.$$

Exercise 28.2 (Smallest singular value). We first observe that

$$\alpha_{\ell^2}^{-1} = \sup_{\mathsf{V}\in\mathbb{C}^I} \frac{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathcal{A}\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}} = \sup_{\mathsf{V}\in\mathbb{C}^I} \frac{\|\mathcal{A}^{-1}\mathcal{A}\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathcal{A}\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}} \le \|\mathcal{A}^{-1}\|_{\ell^2(\mathbb{C}^I)}.$$

Since $\mathbb{C}^I$ is finite-dimensional, there is $\mathsf{V}_* \in \mathbb{C}^I$ s.t. $\|\mathcal{A}^{-1}\|_{\ell^2(\mathbb{C}^I)} = \frac{\|\mathcal{A}^{-1}\mathsf{V}_*\|_{\ell^2(\mathbb{C}^I)}}{\|\mathsf{V}_*\|_{\ell^2(\mathbb{C}^I)}}$. Letting $\mathsf{V}_{**} = \mathcal{A}^{-1}\mathsf{V}_*$, we infer that

$$\alpha_{\ell^2} \le \frac{\|\mathcal{A}\mathsf{V}_{**}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathsf{V}_{**}\|_{\ell^2(\mathbb{C}^I)}} = \frac{\|\mathsf{V}_*\|_{\ell^2(\mathbb{C}^I)}}{\|\mathcal{A}^{-1}\mathsf{V}_*\|_{\ell^2(\mathbb{C}^I)}} = \|\mathcal{A}^{-1}\|_{\ell^2(\mathbb{C}^I)}^{-1}.$$

Exercise 28.3 (ℓ^2-condition number). A direct computation shows that all the components of $\mathcal{Z}\mathsf{X}$ are equal to 2^{1-I}, so that $\|\mathcal{Z}\mathsf{X}\|_{\ell^2(\mathbb{R}^I)} = I^{1/2}2^{1-I}$, and that $\|\mathsf{X}\|_{\ell^2(\mathbb{R}^I)} = (\frac{4}{3}(1-4^{1-I}))^{1/2} \ge (\frac{4}{3})^{1/2}$. Since the last vector of the canonical basis of $\mathbb{R}^I$ is left invariant by $\mathcal{Z}$, we infer that $\|\mathcal{Z}\|_{\ell^2(\mathbb{R}^I)} \ge 1$. This yields

$$\kappa_{\ell^2}(\mathcal{Z}) \ge \|\mathcal{Z}^{-1}\|_{\ell^2(\mathbb{R}^I)} \ge \frac{1}{\|\mathcal{Z}\mathsf{X}\|_{\ell^2(\mathbb{R}^I)}}\|\mathsf{X}\|_{\ell^2(\mathbb{R}^I)} \ge \left(\tfrac{4}{3}\right)^{\frac12} I^{-\frac12}2^{I-1}.$$

If I is large, the matrix $\mathcal{Z}$ is ill-conditioned.

Exercise 28.4 (Local mass matrix, 1D). Setting $h := \frac{1}{I}$, the local mass matrices are, respectively, given by

$$\mathcal{M}^{K,\mathbb{P}_1} = h\begin{pmatrix}\frac13 & \frac16\\ \frac16 & \frac13\end{pmatrix}, \qquad \mathcal{M}^{K,\mathbb{P}_2} = h\begin{pmatrix}\frac{2}{15} & \frac{1}{15} & -\frac{1}{30}\\ \frac{1}{15} & \frac{8}{15} & \frac{1}{15}\\ -\frac{1}{30} & \frac{1}{15} & \frac{2}{15}\end{pmatrix}.$$

Observe the two negative entries in $\mathcal{M}^{K,\mathbb{P}_2}$.

Exercise 28.5 (Stiffness matrix). For the $\mathbb{P}_1$ element, we obtain

$$\nabla\widehat{\lambda}_1 = \begin{pmatrix}1\\0\end{pmatrix}, \quad \nabla\widehat{\lambda}_2 = \begin{pmatrix}0\\1\end{pmatrix}, \quad \nabla\widehat{\lambda}_3 = \begin{pmatrix}-1\\-1\end{pmatrix},$$

so that the stiffness matrix is

$$\widehat{\mathcal{A}} = \begin{pmatrix} \alpha & \delta & \gamma \\ \delta & \alpha & \gamma \\ \gamma & \gamma & \beta \end{pmatrix},$$

with $\alpha := \frac{1}{2}$, $\beta := 1$, $\gamma := -\frac{1}{2}$, and $\delta := 0$. For the $\mathbb{Q}_1$ element, we obtain

$$\nabla\widehat{\theta}_1 = \begin{pmatrix} 1-y \\ -x \end{pmatrix}, \quad \nabla\widehat{\theta}_2 = \begin{pmatrix} y \\ x \end{pmatrix}, \quad \nabla\widehat{\theta}_3 = \begin{pmatrix} -y \\ 1-x \end{pmatrix}, \quad \nabla\widehat{\theta}_4 = \begin{pmatrix} y-1 \\ x-1 \end{pmatrix},$$

so that the stiffness matrix is

$$\widehat{\mathcal{A}} = \begin{pmatrix} a & b & c & b \\ b & a & b & c \\ c & b & a & b \\ b & c & b & a \end{pmatrix},$$

with $a := \frac{2}{3}$, $b := -\frac{1}{6}$, and $c := -\frac{1}{3}$.

Let us consider the domain $D := (0,3)\times(0,2)$. Since the geometric mappings are isometries, we can just combine the entries of the stiffness matrix $\widehat{\mathcal{A}}$. For the first mesh, we obtain

$$\mathcal{A} = \begin{pmatrix} 2\beta+4\alpha & 2\gamma \\ 2\gamma & 2\beta+4\alpha \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}.$$

For the second mesh, we obtain

$$\mathcal{A} = \begin{pmatrix} 4a & 2b \\ 2b & 4a \end{pmatrix} = \begin{pmatrix} \frac{8}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{8}{3} \end{pmatrix}.$$

For the third mesh, we obtain

$$\mathcal{A} = \begin{pmatrix} 2a+\beta+2\alpha & b+\gamma \\ b+\gamma & 2a+\beta+2\alpha \end{pmatrix} = \begin{pmatrix} \frac{10}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{10}{3} \end{pmatrix}.$$

Exercise 28.6 (Sensitivity to perturbations). (i) The estimate results from

$$\|\tilde{\mathsf{X}} - \mathsf{X}\|_{\ell^2(\mathbb{C}^I)} \le \|\mathcal{Z}^{-1}\|_{\ell^2(\mathbb{C}^I)} \|\tilde{\mathsf{B}} - \mathsf{B}\|_{\ell^2(\mathbb{C}^I)},$$
$$\|\mathsf{B}\|_{\ell^2(\mathbb{C}^I)} \le \|\mathcal{Z}\|_{\ell^2(\mathbb{C}^I)} \|\mathsf{X}\|_{\ell^2(\mathbb{C}^I)}.$$

(ii) To prove the second estimate, we observe that $(\check{\mathcal{Z}} - \mathcal{Z})\check{\mathsf{X}} = \mathcal{Z}(\mathsf{X} - \check{\mathsf{X}})$ and infer that

$$\|\check{\mathsf{X}} - \mathsf{X}\|_{\ell^2(\mathbb{C}^I)} \le \|\mathcal{Z}^{-1}\|_{\ell^2(\mathbb{C}^I)} \|\check{\mathcal{Z}} - \mathcal{Z}\|_{\ell^2(\mathbb{C}^I)} \|\check{\mathsf{X}}\|_{\ell^2(\mathbb{C}^I)}.$$

Rearranging the terms proves the assertion.
(iii) Let us prove that the estimate from Step (i) is sharp. Owing to the compactness of the unit ball in finite dimension, there exist $\mathsf{X}_0, \mathsf{B}_0 \in \mathbb{C}^I$ s.t. $\|\mathcal{Z}\mathsf{X}_0\|_{\ell^2(\mathbb{C}^I)} = \|\mathcal{Z}\|_{\ell^2(\mathbb{C}^I)} \|\mathsf{X}_0\|_{\ell^2(\mathbb{C}^I)}$, $\|\mathcal{Z}^{-1}\mathsf{B}_0\|_{\ell^2(\mathbb{C}^I)} = \|\mathcal{Z}^{-1}\|_{\ell^2(\mathbb{C}^I)} \|\mathsf{B}_0\|_{\ell^2(\mathbb{C}^I)}$. This implies that the estimate is sharp. The proof that the estimate from Step (ii) is sharp is similar.

Exercise 28.7 (Stability). Owing to the Cauchy–Schwarz inequality, we infer that

$$\begin{aligned} \alpha_h \|u_h - v_h\|_{V_h} &\le \sup_{w_h \in W_h} \frac{|a_h(u_h - v_h, w_h)|}{\|w_h\|_{W_h}} \\ &= \sup_{\mathsf{W} \in \mathbb{C}^I} \frac{|\mathsf{W}^{\mathsf{H}} \mathcal{A}(\mathsf{U} - \mathsf{V})|}{\|\mathsf{R}_\psi(\mathsf{W})\|_{W_h}} \le \|\mathcal{A}(\mathsf{U} - \mathsf{V})\|_*. \end{aligned}$$

Moreover, we have

$$\begin{aligned}\|\mathsf{B}\|_* &= \sup_{\mathsf{Y}\in\mathbb{C}^I} \frac{|\mathsf{Y}^{\mathsf{H}}\mathsf{B}|}{\|\mathsf{R}_\psi(\mathsf{Y})\|_{W_h}} = \sup_{\mathsf{Y}\in\mathbb{C}^I} \frac{|\ell_h(\mathsf{R}_\varphi(\mathsf{Y}))|}{\|\mathsf{R}_\psi(\mathsf{Y})\|_{W_h}} \\ &= \sup_{\mathsf{Y}\in\mathbb{C}^I} \frac{|a_h(u_h,\mathsf{R}_\varphi(\mathsf{Y}))|}{\|\mathsf{R}_\psi(\mathsf{Y})\|_{W_h}} \le \|a_h\|\|u_h\|_{V_h}.\end{aligned}$$

Combining these two bounds and recalling that $\mathcal{A}\mathsf{U}=\mathsf{B}$ proves the assertion.

Exercise 28.8 (ℓ^∞-norm). (i) Let $\mathsf{V}\in\mathbb{R}^I$ and set $\mathsf{W}:=\mathcal{A}\mathsf{V}$. Since $\mathcal{A}^{-1}\ge 0$, we have

$$\pm\mathsf{V} = \pm\mathcal{A}^{-1}\mathsf{W} \le \|\mathsf{W}\|_{\ell^\infty(\mathbb{R}^I)}\mathcal{A}^{-1}\mathsf{U},$$

where $\mathsf{U}\in\mathbb{R}^I$ has all entries equal to 1. Using the hint leads to

$$\mathcal{A}^{-1}\mathsf{U} \le \frac{1}{\min_{j\in\{1:I\}}(\mathcal{A}\mathsf{Y})_j}\mathsf{Y},$$

so that $\|\mathsf{V}\|_{\ell^\infty(\mathbb{R}^I)} \le \frac{\|\mathsf{Y}\|_{\ell^\infty(\mathbb{R}^I)}}{\min_{j\in\{1:I\}}(\mathcal{A}\mathsf{Y})_j}\|\mathsf{W}\|_{\ell^\infty(\mathbb{R}^I)}$, whence the assertion.
(ii) We observe that $\mathcal{A}$ is a nonsingular M-matrix and that the vector $\mathsf{Y}\in\mathbb{R}^I$ with components $\mathsf{Y}_i := x_i(1-x_i)$ for all $i\in\{1{:}I\}$ is a majorizing vector for $\mathcal{A}$. Indeed $\mathsf{Y}>0$ and $(\mathcal{A}\mathsf{Y})_i = 2h$ for all $i\in\{1{:}I\}$. Using Proposition 28.18 yields $\|\mathcal{A}^{-1}\|_{\ell^\infty(\mathbb{R}^I)} \le \frac{1}{8}h^{-1}$.
(iii) A direct computation shows that

$$\begin{aligned}\mathcal{Z}\mathcal{Z}^{-1} &= (\mathcal{I}+\alpha\mathsf{E}_1\otimes\mathsf{E}_I)(\mathcal{I}-\alpha\mathsf{E}_1\otimes\mathsf{E}_I) \\ &= \mathcal{I}-\alpha^2(\mathsf{E}_1\otimes\mathsf{E}_I)(\mathsf{E}_1\otimes\mathsf{E}_I) \\ &= \mathcal{I}-\alpha^2(\mathsf{E}_1\cdot\mathsf{E}_I)\mathsf{E}_1\otimes\mathsf{E}_I = \mathcal{I}.\end{aligned}$$

Moreover, $\|\mathcal{Z}\|_{\ell^\infty(\mathbb{C}^I)} = \|\mathcal{Z}^{-1}\|_{\ell^\infty(\mathbb{C}^I)} = 1+|\alpha|$, so that $\kappa_{\ell^\infty}(\mathcal{Z}) = (1+|\alpha|)^2$. If α is large, the matrix $\mathcal{Z}$ is ill-conditioned.

Exercise 28.9 (Lumped mass matrix). (i) Consider a cell in the mesh, say $K\in\mathcal{T}_h$. Let $\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_3$ be the three vertices of K and let $\lambda_1,\lambda_2,\lambda_3$ be the associated barycentric coordinates (i.e., the local nodal shape functions). The local mass matrix $\mathcal{M}^K\in\mathbb{R}^{3\times3}$ associated with K is defined to be

$$\mathcal{M}^K_{ij} := \int_K \lambda_i(\boldsymbol{x})\lambda_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = |K|\mathcal{W}_{ij},$$

where the matrix $\mathcal{W}\in\mathbb{R}^{3\times3}$ is given by

$$\mathcal{W} = \begin{bmatrix} \frac{1}{6} & \frac{1}{12} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{6} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{12} & \frac{1}{6} \end{bmatrix}.$$

One way to do this computation is by using the quadrature formula (30.3) (observe that K is indeed a triangle since the mesh is affine).
(ii) The local lumped matrix $\overline{\mathcal{M}}^K\in\mathbb{R}^{3\times3}$ is

$$\overline{\mathcal{M}}^K_{ij} := |K|\overline{\mathcal{W}}_{ij} \qquad \text{with} \qquad \overline{\mathcal{W}} = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}.$$

Of course, since $\overline{\mathcal{W}}$ is diagonal, $\overline{\mathcal{M}}^K$ is diagonal and the assembled matrix $\overline{\mathcal{M}}$ is also diagonal.
(iii) The three eigenvalues of the matrix

$$(\overline{\mathcal{M}}^K)^{-1}(\overline{\mathcal{M}}^K - \mathcal{M}^K) = \overline{\mathcal{W}}^{-1}(\overline{\mathcal{W}} - \mathcal{W}) = 3\begin{bmatrix} \frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} \\ -\frac{1}{12} & \frac{1}{6} & -\frac{1}{12} \\ -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

are $(0, \frac{3}{4}, \frac{3}{4})$.
(iv) Let (Y, λ) be an eigenpair of $\overline{\mathcal{M}}^{-1}(\overline{\mathcal{M}} - \mathcal{M}))$, i.e., $\mathsf{Y}^\mathsf{T}(\overline{\mathcal{M}} - \mathcal{M})\mathsf{Y} = \lambda \mathsf{Y}^\mathsf{T}\overline{\mathcal{M}}\mathsf{Y}$. We infer that

$$\begin{aligned} |\mathsf{Y}^\mathsf{T}(\overline{\mathcal{M}} - \mathcal{M})\mathsf{Y}| &= \Big| \sum_{K\in\mathcal{T}_h} \mathsf{Y}_K^\mathsf{T}(\overline{\mathcal{M}}^K - \mathcal{M}^K)\mathsf{Y}_K \Big| \\ &\le \sum_{K\in\mathcal{T}_h} |K| \|\mathsf{Y}_K\|_{\ell^2} \|\overline{\mathcal{W}} - \mathcal{W}\|_{\ell^2} \|\mathsf{Y}_K\|_{\ell^2}, \end{aligned}$$

where $\mathsf{Y}_K = (\mathsf{Y}_{\mathtt{j\_dof}(K,1)}, \mathsf{Y}_{\mathtt{j\_dof}(K,2)}, \mathsf{Y}_{\mathtt{j\_dof}(K,3)})^\mathsf{T} \in \mathbb{R}^3$ is the vector of the three components of Y that are associated with the vertices of the triangle K ($\mathtt{j\_dof}$ is the connectivity array) and where $\|\cdot\|_{\ell^2}$ denotes either the Euclidian norm or the matrix norm induced by the Euclidean norm. Owing to Step (iii), we infer that $\|\overline{\mathcal{W}} - \mathcal{W}\|_{\ell^2} \le \frac{1}{4}$, which, in turn, implies that

$$|\mathsf{Y}^\mathsf{T}(\overline{\mathcal{M}} - \mathcal{M})\mathsf{Y}| \le \frac{3}{4} \sum_{K\in\mathcal{T}_h} \frac{1}{3}|K| \|\mathsf{Y}_K\|_{\ell^2}^2 = \frac{3}{4} \sum_{K\in\mathcal{T}_h} |K| \mathsf{Y}_K^\mathsf{T} \overline{\mathcal{M}}^K \mathsf{Y}_K = \frac{3}{4} \mathsf{Y}^\mathsf{T} \overline{\mathcal{M}} \mathsf{Y}.$$

In conclusion, we have established that

$$|\mathsf{Y}^\mathsf{T}(\overline{\mathcal{M}} - M)\mathsf{Y}| = |\lambda| \mathsf{Y}^\mathsf{T}\overline{\mathcal{M}}\mathsf{Y} \le \frac{3}{4} \mathsf{Y}^\mathsf{T}\overline{\mathcal{M}}\mathsf{Y},$$

which proves that $\lambda \le \frac{3}{4}$.

Exercise 28.10 (CG). (i) We observe that for all $\mathsf{V} \in \mathbb{R}^I$,

$$\begin{aligned} \frac{1}{2}\|\mathsf{V} - \mathsf{U}\|_\mathcal{A}^2 &= \frac{1}{2}(\mathsf{V} - \mathsf{U})^\mathsf{T}\mathcal{A}(\mathsf{V} - \mathsf{U}) \\ &= \frac{1}{2}\mathsf{V}^\mathsf{T}\mathcal{A}\mathsf{V} - \mathsf{B}^\mathsf{T}\mathsf{V} + \frac{1}{2}\mathsf{U}^\mathsf{T}\mathcal{A}\mathsf{U} \\ &= \mathfrak{J}(\mathsf{V}) + \frac{1}{2}\mathsf{U}^\mathsf{T}\mathcal{A}\mathsf{U}. \end{aligned}$$

This shows that minimizing $\mathfrak{J}$ over $\mathsf{U}_0 + K_m$ is equivalent to minimizing the energy error over this subspace. Proposition 28.20 implies that $\mathsf{V} = \mathsf{U}_m$.
(ii) Since

$$\mathfrak{J}(\mathsf{U}_m + \eta\mathsf{P}_m) = \mathfrak{J}(\mathsf{U}_m) - \eta\mathsf{P}_m^\mathsf{T}\mathsf{R}_m + \frac{1}{2}\eta^2\mathsf{P}_m^\mathsf{T}\mathcal{A}\mathsf{P}_m,$$

we infer that

$$\eta_m = \frac{\mathsf{P}_m^\mathsf{T}\mathsf{R}_m}{\mathsf{P}_m^\mathsf{T}\mathcal{A}\mathsf{P}_m}.$$

From step $(m-1)$ of the CG method, we obtain $\mathsf{P}_m = \mathsf{R}_m + \beta_{m-1}\mathsf{R}_{m-1}$ and since $\mathsf{R}_{m-1}^\mathsf{T}\mathsf{R}_m = 0$ owing to Proposition 28.20, we infer that $\mathsf{P}_m^\mathsf{T}\mathsf{R}_m = \mathsf{R}_m^\mathsf{T}\mathsf{R}_m$, whence we conclude that $\eta_m = \alpha_m$.
(iii) The CG method applied to the preconditioned system $\widetilde{\mathcal{A}}\widetilde{\mathsf{U}} = \widetilde{\mathsf{B}}$ with $\widetilde{\mathcal{A}} := \mathcal{P}_\mathrm{L}^{-1}\mathcal{A}(\mathcal{P}_\mathrm{L}^\mathsf{T})^{-1}$ and $\widetilde{\mathsf{B}} = \mathcal{P}_\mathrm{L}^{-1}\mathsf{B}$ yields iterates $\widetilde{\mathsf{U}}_m$, $\widetilde{\mathsf{P}}_m$, and $\widetilde{\mathsf{R}}_m$ such that $\widetilde{\mathsf{U}}_m = \mathcal{P}_\mathrm{L}^\mathsf{T}\mathsf{U}_m$, $\widetilde{\mathsf{P}}_m = \mathcal{P}_\mathrm{L}^\mathsf{T}\mathsf{P}_m$, and $\widetilde{\mathsf{R}}_m = \mathcal{P}_\mathrm{L}^{-1}\mathsf{R}_m$, where U_m, P_m, and R_m are delivered by Algorithm 28.1.

Algorithm 28.1 Preconditioned CG.

choose U_0, set $\mathsf{R}_0 := \mathsf{B} - \mathcal{A}\mathsf{U}_0$ and $\mathsf{P}_0 := \mathcal{P}^{-1}\mathsf{R}_0$
choose a tolerance `tol` and set $m := 0$
while $\|\mathsf{R}_m\|_{\ell^2} >$ `tol` **do**
  $\alpha_m := \mathsf{R}_m^{\mathsf{T}}\mathcal{P}^{-1}\mathsf{R}_m / \mathsf{P}_m^{\mathsf{T}}\mathcal{A}\mathsf{P}_m$
  $\mathsf{U}_{m+1} := \mathsf{U}_m + \alpha_m \mathsf{P}_m$
  $\mathsf{R}_{m+1} := \mathsf{R}_m - \alpha_m \mathcal{A}\mathsf{P}_m$
  $\beta_m := \mathsf{R}_{m+1}^{\mathsf{T}}\mathcal{P}^{-1}\mathsf{R}_{m+1} / \mathsf{R}_m^{\mathsf{T}}\mathcal{P}^{-1}\mathsf{R}_m$
  $\mathsf{P}_{m+1} := \mathcal{P}^{-1}\mathsf{R}_{m+1} + \beta_m \mathsf{P}_m$
  $m \leftarrow m + 1$
end while

Exercise 28.11 (Complex symmetric system). One readily sees that $\sigma(\mathcal{A}) = \{\mu + \mathsf{i}\sigma \mid \mu \in \sigma(\mathcal{T})\}$, so that $\sigma(\mathcal{A}_*) = \{\mu \pm \mathsf{i}\sigma \mid \mu \in \sigma(\mathcal{T})\}$. If the matrix T is indefinite, the spectrum of $\mathcal{A}_*$ straddles the origin. Furthermore, since $\overline{\mathcal{A}}\mathcal{A} = \mathcal{T}^2 + \sigma^2\mathcal{I}$, $\sigma(\mathcal{A}_{**}) = \{\pm(\mu^2 + \sigma^2)^{\frac{1}{2}} \mid \mu \in \sigma(\mathcal{T})\}$ is included in the real line but straddles the origin with an equal number of eigenvalues on both sides. If one considers the rotated system $-\mathsf{i}\mathcal{A}U = -\mathsf{i}B$ and the first rewriting as a real system, one obtains

$$(-\mathsf{i}\mathcal{A})_* = \begin{pmatrix} \sigma\mathcal{I} & \mathcal{T} \\ -\mathcal{T} & \sigma\mathcal{I} \end{pmatrix},$$

whose spectrum is contained in a line segment parallel to the imaginary line and symmetric with respect to the real line. This is a (much) more favorable situation for Krylov subspace methods.

Chapter 29

Sparse matrices

Exercises

Exercise 29.1 (Retrieving a nonzero entry in CSR format). Write an algorithm to retrieve the value $\mathcal{A}_{ij}$ from the array `aa` stored in CSR format.

Exercise 29.2 (Ellpack (ELL)). Write the arrays needed to store the matrix from Example 29.3 in the Ellpack format. Write an algorithm that performs a matrix-vector multiplication in this format.

Exercise 29.3 (Coordinate format (COO)). Let $\mathcal{A}$ be a $I \times I$ sparse matrix. Consider the storage format where one stores the nonzero entries $\mathcal{A}_{ij}$ in the array `aa`(1:nnz) and stores in the same order the row and columns indices in the integer arrays `ia`(1:nnz) and `ja`(1:nnz), respectively. (i) Use this format to store the matrix defined in (29.4). (ii) Write an algorithm to perform a matrix-vector product in this format. Compare with the CSR format.

Exercise 29.4 (Storage). Consider the storage format for sparse $I \times I$ matrices where one stores the nonzero entries $\mathcal{A}_{ij}$ in the array `aa`(1:nnz) and stores in the same order the integer $(i-1)I + j$ in the integer array `ja`(1:nnz). (i) Use this format for the matrix defined in (29.4). (ii) Write an algorithm to do matrix-vector products in this format. Compare with the CSR format.

Exercise 29.5 (Greedy coloring). (i) Prove that the total number of colors found by Algorithm 29.5 is at most equal to 1 plus the largest degree in the graph. (ii) Assume that a graph G can be colored with two colors only. Prove that if the BFS reordering is used to initialize `traverse`, then Algorithm 29.5 finds a two-color partitioning. (*Hint*: by induction on the number of level sets.)

Exercise 29.6 (Multicolor ordering). Prove Proposition 29.10.

Exercise 29.7 (CMK reordering). Give the sparsity pattern and the CMK reordering for the matrix shown in Figure 29.4.

Solution to exercises

Exercise 29.1 (Retrieving a nonzero entry in CSR format). Algorithm 29.1 shows how to retrieve the value of $\mathcal{A}_{ij}$ from the array `aa` assuming that the entry $\mathcal{A}_{ij}$ is nonzero:

Algorithm 29.1 Retrieving $\mathcal{A}_{ij} \neq 0$ in CSR format.

```
for p ∈ {ia(i): ia(i+1)−1} do
  if ja(p) := j then
    value := aa(p); Exit loop over p
  end if
end for
```

Exercise 29.2 (Ellpack (ELL)). For the 5×5 matrix shown in (29.4), $N_{\text{row}} = 4$ and

$$\texttt{aa} = \begin{bmatrix} 1. & 2. & 0. & 0. \\ 3. & 4. & 5. & 0. \\ 6. & 7. & 8. & 9. \\ 10. & 11. & 0. & 0. \\ 12. & 0. & 0. & 0. \end{bmatrix}, \qquad \texttt{ja} = \begin{bmatrix} 1 & 4 & 4 & 4 \\ 1 & 2 & 4 & 4 \\ 1 & 3 & 4 & 5 \\ 3 & 4 & 4 & 4 \\ 5 & 5 & 5 & 5 \end{bmatrix}.$$

The following algorithm shows how to evaluate the matrix-vector multiplication $y = \mathcal{A}x$ in the Ellpack format.

Algorithm 29.2 Matrix-vector multiplication in Ellpack format.

```
for i ∈ {1:I} do;  yi := 0
  for p ∈ {1:N_row} do
    yi := yi + aa(i,p) * x(ja(i,p))
  end for
  y(i) := yi
end for
```

Exercise 29.3 (Coordinate format (COO)). (i) One possibility could be

$$\begin{aligned} \texttt{aa} &= [1.\ 2.\ 3.\ 4.\ 5.\ 6.\ 7.\ 8.\ 9.\ 10.\ 11.\ 12.] \\ \texttt{ia} &= [1\ 1\ 2\ 2\ 2\ 3\ 3\ 3\ 3\ 4\ 4\ 5] \\ \texttt{ja} &= [1\ 4\ 1\ 2\ 4\ 1\ 3\ 4\ 5\ 3\ 4\ 5] \end{aligned}$$

Another one could be

$$\begin{aligned} \texttt{aa} &= [1.\ 3.\ 6.\ 4.\ 7.\ 10.\ 2.\ 5.\ 8.\ 11.\ 9.\ 12.] \\ \texttt{ia} &= [1\ 2\ 3\ 2\ 3\ 4\ 1\ 2\ 3\ 4\ 3\ 5] \\ \texttt{ja} &= [1\ 1\ 1\ 2\ 3\ 3\ 4\ 4\ 4\ 4\ 5\ 5] \end{aligned}$$

(ii) We now write an algorithm for the matrix-vector multiplication in the coordinate format. The algorithm essentially consists of a single loop, whereas there are two nested loops for the CSR format.

Algorithm 29.3 Matrix-vector multiplication in coordinate format.

for $i \in \{1{:}I\}$ **do**
  $y(i) := 0$
end for
for $p \in \{1{:}nnz\}$ **do**
  $y(\texttt{ia}(p)) := y(\texttt{ia}(p)) + \texttt{aa}(p) * x(\texttt{ja}(p))$
end for

Exercise 29.4 (Storage). One possibility could be

$$\texttt{aa} = [1.\ 2.\ 3.\ 4.\ 5.\ 6.\ 7.\ 8.\ 9.\ 10.\ 11.\ 12.]$$
$$\texttt{ja} = [1\ 4\ 6\ 7\ 9\ 11\ 13\ 14\ 15\ 18\ 19\ 25]$$

Another one could be

$$\texttt{aa} = [1.\ 3.\ 6.\ 4.\ 7.\ 10.\ 2.\ 5.\ 8.\ 11.\ 9.\ 12.]$$
$$\texttt{ja} = [1\ 2\ 3\ 7\ 13\ 14\ 16\ 17\ 18\ 19\ 23\ 25]$$

(ii) We now write an algorithm for the matrix-vector multiplication in the proposed format. The algorithm essentially consists of a single loop, whereas there are two nested loops for the CSR format.

Algorithm 29.4 Matrix-vector multiplication.

for $i \in \{1{:}I\}$ **do**
  $y(i) := 0$
end for
for $p \in \{1{:}nnz\}$ **do**
  $j = \text{modulo}(\texttt{ja}(p) - 1, I) + 1$
  $i = (\texttt{ja}(p) - j)/I + 1$
  $y(i) := y(i) + \texttt{aa}(p) * x(j)$
end for

Exercise 29.5 (Greedy coloring). (i) Let $k \geq 0$ be the largest degree in the graph. Let $j \in \{1{:}I\}$. Assume that $\text{Adj}(j) \neq \emptyset$ (otherwise the greedy coloring algorithm (Algorithm 29.5) gives $\texttt{color}(j) = 1 \leq k+1$). Assume that $\min\{l > 0 \mid l \notin \texttt{color}(\text{Adj}(j))\} \geq k+2$. This means that $\{1, \ldots, k+1\} \subset \texttt{color}(\text{Adj}(j))$, which in turn implies that the cardinality of $\texttt{color}(\text{Adj}(j))$ is at least $k+1$. Hence, the cardinality of $\text{Adj}(j)$ is at least $k+1$, which is in contradiction with the definition of k. We then conclude that $\min\{l > 0 \mid l \notin \texttt{color}(\text{Adj}(j))\} \leq k+1$, and this implies that $\texttt{color}(j)$ as defined by the greedy coloring algorithm is less than $k+1$. This proves that the total number of colors found by the algorithm is at most $k+1$.
(ii) Since we know that the graph can be colored with two colors only, there exists a (theoretical) graph coloring map $\texttt{color}^{\texttt{th}} : V \to \{1, 2\}$. Let $\{L_l\}_{1 \in \{1:l\}}$ be a set of level sets of the graph G.

Assume that `traverse` is based on the BFS reordering using these level sets. Let us prove by induction that $\text{card}(\texttt{color}^{\texttt{th}}(L_l)) = 1$ and that the greedy coloring algorithm (Algorithm 29.5) gives the same color, $\text{modulo}(l-1, 2)+1$, to all the vertices in the same level set L_l. The induction hypothesis holds true for $l = 1$ since the first level set L_1 has only one vertex. Assume that $k \geq 2$, otherwise there is nothing to prove. Let $l \geq 1$ and $j, n \in L_{l+1}$. Since `traverse` is based on the

BFS reordering, one of the neighboring vertices of j, say $i(j)$, must belong to the level set L_l. The same argument shows that one of the neighboring vertices of n, say $i(n)$, must belong to the level set L_l. But the graph can be colored with two colors only. This means that two neighboring vertices must have different colors. As a result, one must have

$$\begin{aligned}\texttt{color}^{\texttt{th}}(j) &= \text{modulo}(\texttt{color}^{\texttt{th}}(i(j),2)+1\\ &= \text{modulo}(\texttt{color}^{\texttt{th}}(i(n),2)+1 = \texttt{color}^{\texttt{th}}(n).\end{aligned}$$

Hence, we have proved that $\text{card}(\texttt{color}^{\texttt{th}}(L_{l+1})) = 1$. This argument also shows that j cannot have any neighbor in L_{l+1}. Hence, by construction of the level sets, the neighbors of j can only belong to $L_l \cap L_{l+2}$. Since the vertices in L_{l+2} (if $l+2 \geq k$) have not been visited yet, their color assigned by Algorithm 29.5 is zero. Hence, we have

$$\begin{aligned}\min\{s \geq 1 \mid s \notin \texttt{color}(\text{Adj}(j))\} &= \min\{s \geq 1 \mid s \neq \texttt{color}(i(j))\}\\ &= \min\{s \geq 1 \mid s \neq \text{modulo}(l-1,2)+1\}\\ &= \text{modulo}(l,2)+1.\end{aligned}$$

In other words, $\texttt{color}(j) = \text{modulo}(l,2)+1$ for every $j \in L_{l+1}$. This proves the assertion.

Exercise 29.6 (Multicolor ordering). Let $\mathcal{B}$ be the reordered matrix. We define a $\texttt{k\_max}\times\texttt{k\_max}$ block structure of $\mathcal{B}$ by saying that $\mathcal{B}_{ij}$ is in the block $k\times l$ if $\texttt{color}(i) = k$ and $\texttt{color}(j) = l$. Let $i, j \in \{1{:}I\}$ be s.t. $\mathcal{B}_{ij}$ is in the k-th diagonal block. This means that i and j have the same color k. Assume that $i \neq j$. Then $j \notin \text{Adj}(i)$, otherwise j and i would have different colors. This means that $\mathcal{B}_{ij} = 0$ (recall that $j \in \text{Adj}(i)$ iff $\mathcal{B}_{ij} \neq 0$ or $\mathcal{B}_{ji} \neq 0$). This proves that the k-th diagonal block of $\mathcal{B}$ is diagonal.

Exercise 29.7 (CMK reordering). Starting from the vertex 8, the level sets are

$$L_1 = \{8\}, \quad L_2 = \{2\}, \quad L_3 = \{1,3,4,5\}, \quad L_4 = \{6,7\}.$$

One possibility for the permutation index corresponding to the CMK reordering is

$$\texttt{perm} = \{8,2,1,5,3,4,6,7\}.$$

The sparsity pattern and adjacency graph are shown in Figure 29.1.

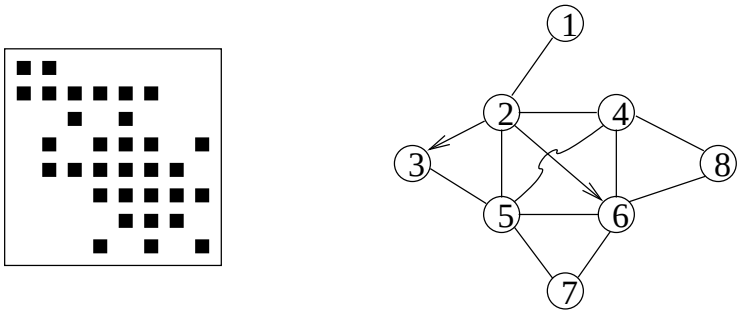


Figure 29.1: Sparsity pattern (left) and adjacency graph (right) of a 8×8 sparse matrix.

Chapter 30

Quadratures

Exercises

Exercise 30.1 (Quadratures on simplices). Let K be a simplex in $\mathbb{R}^d$. Let $\boldsymbol{z}_K$ be the barycenter of K, let $\{\boldsymbol{z}_i\}_{i\in\{0:d\}}$ be the vertices of K, and let $\{\boldsymbol{m}_i\}_{i\in\{0:d\}}$ be the midpoints of the edges of K. Consider the following quadratures: $\{\boldsymbol{z}_K\}$, $\{|K|\}$; $\{\boldsymbol{z}_i\}_{i\in\{0:d\}}$, $\{\frac{1}{d+1}|K|\}$; $\{\boldsymbol{m}_i\}_{i\in\{0:d\}}$, $\{\frac{1}{d+1}|K|\}$. (i) Prove that the first and the second quadratures are of order one. (ii) Prove that the third one is of order two for $d=2$.

Exercise 30.2 (Quadrature for $\mathbb{Q}_{2,d}$). Let $\widehat{K}:=[0,1]^d$ be the unit hypercube. Let $\widehat{\boldsymbol{a}}_{i_1\dots i_d}:=(\frac{i_1}{2},\dots,\frac{i_d}{2})$, $i_1,\dots,i_d\in\{0{:}2\}$. Show that the quadrature $\int_{\widehat{K}} f(\widehat{\boldsymbol{x}})\,\mathrm{d}\widehat{x}\approx\sum_{i_1,\dots i_d} w_{i_1\dots i_d} f(\widehat{\boldsymbol{a}}_{i_1\dots i_d})$ where $w_{i_1\dots i_d}:=\frac{1}{6^d}\prod_{k=1}^d(3i_k(2-i_k)+1)$ is exact for all $f\in\mathbb{Q}_{2,d}$. (*Hint*: write the $\mathbb{Q}_{2,d}$ Lagrange shape functions in tensor-product form and use Simpson's rule in each direction.)

Exercise 30.3 (Global quadrature error). Prove that

$$\left|\int_D \phi(\boldsymbol{x})\,\mathrm{d}x-\sum_{K\in\mathcal{T}_h}\sum_{l\in\{1:l_{\mathcal{Q}}\}}\omega_{lK}\,\phi(\boldsymbol{\xi}_{lK})\right|\le ch^m|D|^{1-\frac{1}{p}}|\phi|_{W^{m,p}(D)},$$

for all $\phi\in W^{m,p}(D)$ and all $h\in\mathcal{H}$. (*Hint*: use Lemma 30.9.)

Exercise 30.4 (Quadrature error with polynomial). The goal is to prove (30.7). We are going to make use of (30.6) formulated as follows: $|E_K(\psi q)|\le c\,h_K^\mu|\psi|_{W^{\mu,\infty}(K)}\|q\|_{L^1(K)}$ for all $q\in\mathbb{P}_{\nu,d}\circ\boldsymbol{T}_K$ where $\mu+\nu-1\le k_{\mathcal{Q}}$, $\mu,\nu\in\mathbb{N}$. (i) Prove that $|E_K(\phi\underline{p}_K)|\le ch_K^m|\phi|_{W^{m,\infty}(K)}\|p\|_{L^1(K)}$, where $\underline{p}_K$ is the mean value of p over K. (ii) Prove (30.7). (*Hint*: use Step (i) with $\mu:=m-1$.)

Exercise 30.5 (Surface quadrature). Assume $d=3$. Let F be a face of a mesh cell. Let $\widehat{F}\subset\mathbb{R}^2$ be a reference face and let $\boldsymbol{T}_F:\widehat{F}\to F$ be the geometric mapping for F. Let $\boldsymbol{t}_1(\widehat{\boldsymbol{s}}),\boldsymbol{t}_2(\widehat{\boldsymbol{s}})$ be the two column vectors of the Jacobian matrix of $\boldsymbol{T}_F(\widehat{\boldsymbol{s}})$, say $\mathbb{J}_F(\widehat{\boldsymbol{s}}):=[\boldsymbol{t}_1(\widehat{\boldsymbol{s}}),\boldsymbol{t}_2(\widehat{\boldsymbol{s}})]\in\mathbb{R}^{3\times2}$. (i) Compute the metric tensor $\mathrm{g}_F:=\mathbb{J}_F^{\mathsf{T}}\mathbb{J}_F\in\mathbb{R}^{2\times2}$ in terms of the dot products $\boldsymbol{t}_i\cdot\boldsymbol{t}_j$, $i,j\in\{1,2\}$. (ii) Show that $\mathrm{d}s=\|\boldsymbol{t}_1(\widehat{\boldsymbol{s}})\times\boldsymbol{t}_2(\widehat{\boldsymbol{s}})\|_{\ell^2(\mathbb{R}^3)}\,\mathrm{d}\widehat{\boldsymbol{s}}$. (*Hint*: use Lagrange's identity, that is, $\|\boldsymbol{a}\|_{\ell^2(\mathbb{R}^3)}^2\|\boldsymbol{b}\|_{\ell^2(\mathbb{R}^3)}^2-(\boldsymbol{a}\cdot\boldsymbol{b})^2=\|\boldsymbol{a}\times\boldsymbol{b}\|_{\ell^2(\mathbb{R}^3)}^2$ for any pair of vectors $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^3$, and recall that $\mathrm{d}s=\sqrt{\det(\mathrm{g}_F)}\,\mathrm{d}\widehat{s}$). (iii) Given a quadrature $\{\widehat{\boldsymbol{s}}_l,\widehat{w}_l\}_{l\in\{1:l_{\mathcal{Q}}^\partial\}}$ on $\widehat{F}$, generate the quadrature on F.

Exercise 30.6 (Asssembling). Let $D := (0,1)^2$. Consider the problem $-\Delta u + u = 1$ in D and $u_{|\partial D} = 0$. (i) Approximate its solution with $\mathbb{P}_1$ H^1-conforming finite elements on the two meshes shown in Figure 30.1. (ii) Evaluate the discrete solution in both cases. (*Hint*: there is only one degree of freedom in both cases, see Exercise 28.5 for computing the gradient part of the stiffness coefficient and use a quadrature from Table 30.1 for the zero-order term.) (iii) For a fine mesh composed of 800 elements, we have $u_h(\frac{1}{2},\frac{1}{2}) \approx 0.0702$. Comment.

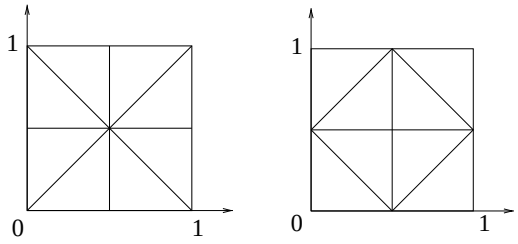


Figure 30.1: Illustration for Exercise 30.6.

Exercise 30.7 (Discrete data). Adapt Algorithm 30.1 to the case where $(\mathrm{d}_{k_1k_2})_{k_1,k_2\in\{1:d\}}$, $(\boldsymbol{\beta}_{k_1})_{k_1\in\{1:d\}}$, and μ are known in the discrete space V_h. (*Hint*: let `dif`, `beta`, and `mu` be the corresponding coordinate vectors, and observe that $\mu(\boldsymbol{\xi}_{lK_m}) = \sum_{n\in\{1:n_{\mathrm{sh}}\}} \mathtt{mu}(\mathtt{j\_dof}(m,i))\times\mathtt{theta}(n,l)$, etc.)

Exercise 30.8 (Assembling of RHS). Write the assembling algorithm for the right-hand side vector in the case where $F(\boldsymbol{\xi}, w_h) := f(\boldsymbol{\xi})w_h(\boldsymbol{\xi}) + \sum_{k_1\in\{1:d\}} \boldsymbol{\beta}_{k_1}(\boldsymbol{\xi})\frac{\partial w_h}{\partial x_{k_1}}(\boldsymbol{\xi})$ with analytically known data.

Solution to exercises

Exercise 30.1 (Quadratures on simplices). (i) Consider the first quadrature. Let $i \in \{0{:}d\}$ and λ_i be the i-th barycentric coordinate. We have $\int_K \lambda_i \,\mathrm{d}x = \frac{1}{(d+1)}|K|$ and $\lambda_i(\boldsymbol{z}_K) = \frac{1}{d+1}$. Hence, $\int_K \lambda_i \,\mathrm{d}x = \lambda_i(\boldsymbol{z}_K)|K|$. This proves that the first quadrature is at least of order 1 since $\mathbb{P}_{1,d} = \mathrm{span}\{\lambda_i\}_{i\in\{0:d\}}$. To show that the quadrature is not of order 2, we observe that $\int_K \lambda_i^2 \,\mathrm{d}x = |K|\frac{2}{(d+1)(d+2)}$, whereas the quadrature gives $|K|\frac{1}{(d+1)^2}$. Let us consider the second quadrature. We have $\int_K \lambda_i \,\mathrm{d}x = \frac{1}{(d+1)}|K|\sum_{j\in\{0:d\}} \lambda_i(\boldsymbol{z}_j)$ since $\sum_{j\in\{0:d\}} \lambda_i(\boldsymbol{z}_j) = 1$. This proves that the second quadrature is at least of order 1 since $\mathbb{P}_{1,d} = \mathrm{span}\{\lambda_i\}_{i\in\{0:d\}}$. To show that the quadrature is not of order 2, we observe again that $\int_K \lambda_i^2 \,\mathrm{d}x = |K|\frac{2}{(d+1)(d+2)}$, whereas the quadrature gives $|K|\frac{1}{d+1}$. (ii) Let us now consider the third quadrature for $d = 2$. We have $\int_K \lambda_i \,\mathrm{d}x = \frac{1}{(d+1)}|K| = \frac{1}{(d+1)}|K|\sum_{j\in\{0:2\}} \lambda_i(\boldsymbol{m}_j)$, where we used that $\sum_{j\in\{0:2\}} \lambda_i(\boldsymbol{m}_j) = 1$ in $\mathbb{R}^2$. Since $\int_K \lambda_i\lambda_j \,\mathrm{d}x = \frac{1}{(d+2)(d+1)}|K|$ with $i \neq j$ (see (30.3)), we infer that $\int_K \lambda_i\lambda_j \,\mathrm{d}x = \frac{1}{(d+1)}|K|\sum_{k\in\{0:2\}} \lambda_i(\boldsymbol{m}_k)\lambda_j(\boldsymbol{m}_k)$ since we have $\sum_{k\in\{0:2\}} \lambda_i(\boldsymbol{m}_k)\lambda_j(\boldsymbol{m}_k) = \frac{1}{4} = \frac{1}{d+2}$. This proves that the third quadrature is at least of order 2 since $\mathbb{P}_{2,2} = \mathrm{span}\{\lambda_0, \lambda_1, \lambda_2, \lambda_0\lambda_1, \lambda_0\lambda_2, \lambda_1\lambda_2\}$. To show that the quadrature is not of order 3, we observe that $\int_K \lambda_i^3 \,\mathrm{d}x = |K|\frac{6}{(d+1)(d+2)(d+3)}$, whereas the quadrature gives $|K|\frac{d}{8(d+1)}$.

Exercise 30.2 (Quadrature for $\mathbb{Q}_{2,d}$). Let $\widehat{\theta}_{i_1\dots i_d}$ be the $\mathbb{Q}_{2,d}$ Lagrange shape function associated with the node $\widehat{\boldsymbol{a}}_{i_1\dots i_d} = (\frac{i_1}{2},\dots,\frac{i_d}{2})$, $i_1,\dots,i_d \in \{0{:}2\}$. This shape function is s.t. $\widehat{\theta}_{i_1\dots i_d}(\widehat{\boldsymbol{x}}) =$

$\widehat{\theta}_{i_1}(\widehat{x}_1)\dots\widehat{\theta}_{i_d}(\widehat{x}_d)$ where $\widehat{\boldsymbol{x}} := (\widehat{x}_1,\dots,\widehat{x}_d)^\mathsf{T}$ and $\{\widehat{\theta}_i\}_{i\in\{0:2\}}$ are the univariate $\mathbb{Q}_{2,d}$ Lagrange basis functions associated with the nodes 0, $\frac{1}{2}$, and 1. Using Simpson's rule yields

$$\begin{aligned}\int_{\widehat{K}}\widehat{\theta}_{i_1\dots i_d}(\widehat{\boldsymbol{x}})\,\mathrm{d}\widehat{x} &= \prod_{k\in\{1:d\}}\left(\int_0^1\widehat{\theta}_{i_k}(\widehat{x}_k)\,\mathrm{d}\widehat{x}_k\right)\\ &= \prod_{k\in\{1:d\}}\frac{1}{6}(\widehat{\theta}_{i_k}(0)+4\widehat{\theta}_{i_k}(\tfrac{1}{2})+\widehat{\theta}_{i_k}(1))\\ &= \frac{1}{6^d}\prod_{k\in\{1:d\}}\left(\sum_{l\in\{0:2\}}(3l(2-l)+1)\widehat{\theta}_{i_k}(\tfrac{l}{2})\right)\\ &= \frac{1}{6^d}\prod_{k\in\{1:d\}}(3i_k(2-i_k)+1) = w_{i_1\dots i_d}.\end{aligned}$$

The conclusion follows readily since $(\widehat{\theta}_{i_1\dots i_d})_{i_1,\dots,i_d\in\{0:2\}}$ is a basis of $\mathbb{Q}_{2,d}$.

Exercise 30.3 (Global quadrature error). Owing to Lemma 30.9, we infer that

$$\begin{aligned}\left|\int_D\phi(\boldsymbol{x})\,\mathrm{d}x-\sum_{K\in\mathcal{T}_h}\sum_{l\in\{1:l_\mathcal{Q}\}}\omega_{lK}\,\phi(\boldsymbol{\xi}_{lK})\right| &\le \sum_{K\in\mathcal{T}_h}|E_K(\phi)|\\ &\le c\sum_{K\in\mathcal{T}_h}h_K^m|K|^{1-\frac{1}{p}}|\phi|_{W^{m,p}(K)}\\ &\le c\,h^m\Big(\sum_{K\in\mathcal{T}_h}|K|\Big)^{\frac{1}{p'}}\Big(\sum_{K\in\mathcal{T}_h}|\phi|^p_{W^{m,p}(K)}\Big)^{\frac{1}{p}},\end{aligned}$$

with $\frac{1}{p}+\frac{1}{p'}=1$, where we used Hölder's inequality in $\mathbb{R}^{N_\mathrm{c}}$ (where N_c denotes the number of mesh cells). The conclusion follows from $\sum_{K\in\mathcal{T}_h}|K| = |D|$.

Exercise 30.4 (Quadrature error with polynomial). (i) We use the hint with $\psi := \phi$, $q := \underline{p}_K$, $\mu := m$, and $\nu := 0$. This is legitimate since $1\le n$ implies that $\mu+\nu-1 = m-1 \le m+n-2\le k_\mathcal{Q}$. Hence, $|E_K(\phi\underline{p}_K)|\le ch_K^m|\phi|_{W^{m,\infty}(K)}\|\underline{p}_K\|_{L^1(K)}$. We conclude by observing that $\|\underline{p}_K\|_{L^1(K)}\le\|p\|_{L^1(K)}$.
(ii) We apply again the hint with $\mu := m-1$ and $\nu = n$. Notice that $\mu\ge 0$ since $m\ge 1$ and that $\mu+\nu-1 = m+n-2\le k_\mathcal{Q}$. This yields

$$|E_K(\phi(p-\underline{p}_K))|\le c\,h_K^{m-1}|\phi|_{W^{m-1,\infty}(K)}\|p-\underline{p}_K\|_{L^1(K)},$$

and $\|p-\underline{p}_K\|_{L^1(K)}\le ch_K\|\nabla p\|_{\boldsymbol{L}^1(K)}$ follows from the Poincaré–Steklov inequality (see (3.8) or (12.13)). Since $E_K(\phi p) = E_K(\phi(p-\underline{p}_K))+E_K(\phi\underline{p}_K)$, we conclude by using the bound from Step (i).

Exercise 30.5 (Surface quadrature). (i) By definition, we have

$$\mathrm{g}_F := \mathbb{J}_F^\mathsf{T}\mathbb{J}_F := \begin{bmatrix}\boldsymbol{t}_1^\mathsf{T}\\ \boldsymbol{t}_2^\mathsf{T}\end{bmatrix}\begin{bmatrix}\boldsymbol{t}_1,\boldsymbol{t}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{t}_1\cdot\boldsymbol{t}_1 & \boldsymbol{t}_1\cdot\boldsymbol{t}_2\\ \boldsymbol{t}_2\cdot\boldsymbol{t}_1 & \boldsymbol{t}_2\cdot\boldsymbol{t}_2\end{bmatrix}.$$

(ii) We have $\mathrm{d}s = \sqrt{\det(\mathrm{g}_F)}\,\mathrm{d}\widehat{s}$, but Lagrange's identity gives

$$\det(\mathrm{g}_F) = \|\boldsymbol{t}_1\|^2_{\ell^2(\mathbb{R}^3)}\|\boldsymbol{t}_2\|^2_{\ell^2(\mathbb{R}^3)}-(\boldsymbol{t}_1\cdot\boldsymbol{t}_2)^2 = \|\boldsymbol{t}_1\times\boldsymbol{t}_2\|^2_{\ell^2(\mathbb{R}^3)},$$

whence $\mathrm{d}s = \|\boldsymbol{t}_1(\widehat{\boldsymbol{s}})\times\boldsymbol{t}_2(\widehat{\boldsymbol{s}})\|_{\ell^2(\mathbb{R}^3)}\,\mathrm{d}\widehat{s}$.
(iii) Let $\{\widehat{\boldsymbol{s}}_l,\widehat{w}_l\}_{l\in\{1:l^\partial_{\mathcal{Q}}\}}$ be a quadrature on $\widehat{F}$. We have

$$\begin{aligned}\int_F \phi(\boldsymbol{x})\,\mathrm{d}s &= \int_{\widehat{F}} \phi(\boldsymbol{T}_F(\widehat{\boldsymbol{s}}))\|\boldsymbol{t}_1(\widehat{\boldsymbol{s}})\times\boldsymbol{t}_2(\widehat{\boldsymbol{s}})\|_{\ell^2(\mathbb{R}^3)}\,\mathrm{d}\widehat{s}\\ &\approx \sum_{l\in\in\{1:l^\partial_{\mathcal{Q}}\}} \phi(\boldsymbol{T}_F(\widehat{\boldsymbol{s}}_l))\widehat{w}_l\|\boldsymbol{t}_1(\widehat{\boldsymbol{s}}_l)\times\boldsymbol{t}_2(\widehat{\boldsymbol{s}}_l)\|_{\ell^2(\mathbb{R}^3)}.\end{aligned}$$

The quadrature on F is $\{\boldsymbol{T}_F(\widehat{\boldsymbol{s}}_l),\widehat{w}_l\|\boldsymbol{t}_1(\widehat{\boldsymbol{s}}_l)\times\boldsymbol{t}_2(\widehat{\boldsymbol{s}}_l)\|_{\ell^2(\mathbb{R}^3)}\}_{l\in\{1:l^\partial_{\mathcal{Q}}\}}$.

Exercise 30.6 (Asssembling). (i) Let us consider a reference triangle $\widehat{K}$ with vertices $(1,0)$, $(0,1)$, $(0,0)$ and let $\widehat{\lambda}_1,\widehat{\lambda}_2,\widehat{\lambda}_3$ be the corresponding barycentric coordinates. The local stiffness matrix has been computed in Exercise 28.5. Using the same enumeration convention as in Exercise 28.5, we have

$$\Big(\int_{\widehat{K}} \nabla\widehat{\lambda}_m\cdot\nabla\widehat{\lambda}_n\,\mathrm{d}\widehat{x}\Big)_{m,n\in\{1:3\}} = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2}\\ 0 & \frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & -\frac{1}{2} & 1\end{pmatrix} =: \mathcal{D}.$$

For the zero-order term, we can use the quadrature of degree 2 from Table 30.1 based on the three edge midpoints. We obtain

$$\Big(\int_{\widehat{K}} \widehat{\lambda}_m\widehat{\lambda}_n\,\mathrm{d}\widehat{x}\Big)_{m,n\in\{1:3\}} = \begin{pmatrix} \frac{1}{12} & \frac{1}{24} & \frac{1}{24}\\ \frac{1}{24} & \frac{1}{12} & \frac{1}{24}\\ \frac{1}{24} & \frac{1}{24} & \frac{1}{12}\end{pmatrix} =: \mathcal{M}.$$

The stiffness matrix is then $\mathcal{A} := \mathcal{D}+\mathcal{M}$. For the assembly procedure, we have $h=\frac{1}{2}$, $|K| = \frac{h^2}{2} = \frac{1}{8}$, $|\widehat{K}| = \frac{1}{2}$, and

$$\begin{aligned}\int_K \nabla\varphi_i\cdot\nabla\varphi_j\,\mathrm{d}x &= h^{-2}\frac{|K|}{|\widehat{K}|}\int_{\widehat{K}} \nabla\widehat{\theta}_m\cdot\nabla\widehat{\theta}_n\,\mathrm{d}\widehat{x} = \int_{\widehat{K}} \nabla\widehat{\theta}_m\cdot\nabla\widehat{\theta}_n\,\mathrm{d}\widehat{x},\\ \int_K \varphi_i\varphi_j\,\mathrm{d}x &= \frac{|K|}{|\widehat{K}|}\int_{\widehat{K}} \widehat{\theta}_m\widehat{\theta}_n\,\mathrm{d}\widehat{x} = \frac{1}{4}\int_{\widehat{K}} \widehat{\theta}_m\widehat{\theta}_n\,\mathrm{d}\widehat{x},\\ \int_K \varphi_i\,\mathrm{d}x &= \frac{|K|}{|\widehat{K}|}\int_{\widehat{K}} \widehat{\theta}_m\,\mathrm{d}\widehat{x} = \frac{1}{4}\int_{\widehat{K}} \widehat{\theta}_m\,\mathrm{d}\widehat{x} = \frac{1}{24},\end{aligned}$$

with $i := \mathtt{j\_dof}(K,m)$ and $j := \mathtt{j\_dof}(K,n)$. Here, we have only one global shape function so that $i:=1$ and $j:=1$.

For the mesh on the left, we obtain for the stiffness coefficient and the right-hand side $\mathcal{A}_{11} = 4(\mathcal{D}_{11}+\mathcal{D}_{22}) + 4\frac{1}{4}(\mathcal{M}_{11}+\mathcal{M}_{22}) = 4+\frac{1}{6} = \frac{25}{6}$ and $\mathsf{F}_1 = 8\frac{1}{24} = \frac{1}{3}$, respectively, so that the approximate solution is $\mathsf{U} = \frac{2}{25}\approx 0.08$.

For the mesh on the right, we obtain $\mathcal{A}_{11} = 4\mathcal{D}_{33} + 4\frac{1}{4}\mathcal{M}_{33} = 4+\frac{1}{12} = \frac{49}{12}$ and $\mathsf{F}_1 = 4\frac{1}{24} = \frac{1}{6}$, so that $\mathsf{U} = \frac{2}{49}\approx 0.04$.

We observe that the first mesh leads to a more accurate solution. The advantage of this mesh is that all the mesh cells contribute to the matrix and the right-hand side vector. In the second mesh, the four triangles having two boundary edges do not contribute to the approximation.

Exercise 30.7 (Discrete data). We use the hint to compute the values of all the coefficients at the Gauss nodes on every mesh cell. The assembling is done in Algorithm 30.1.

Exercise 30.8 (Assembling of RHS). The assembling is done in Algorithm 30.2.

Algorithm 30.1 Assembling of $\mathcal{A}_{\mathcal{Q}}$ for discrete data.

$\mathcal{A}_{\mathcal{Q}} = 0$
for $m \in \{1{:}N_{\mathrm{c}}\}$ **do**
  for $l \in \{1{:}l_{\mathcal{Q}}\}$ **do**; $\mathtt{tmp} := 0$
    for $k_1 \in \{1{:}d\}$ **do**
      for $k_2 \in \{1{:}d\}$ **do**
        $\mathtt{dif\_l}(k_1,k_2) := \sum_{n\in\{1:n_{\mathrm{sh}}\}} \mathtt{dif}(k_1,k_2,\mathtt{j\_dof}(m,n)) * \mathtt{theta}(n,l)$
      end for
      $\mathtt{beta\_l}(k_1) := \sum_{n\in\{1:n_{\mathrm{sh}}\}} \mathtt{beta}(k_1,\mathtt{j\_dof}(m,n)) * \mathtt{theta}(n,l)$
    end for
    $\mathtt{mu\_l} := \sum_{n\in\{1:n_{\mathrm{sh}}\}} \mathtt{mu}(\mathtt{j\_dof}(m,n)) * \mathtt{theta}(n,l)$
    for $ni \in \{1{:}n_{\mathrm{sh}}\}$ **do**
      for $nj \in \{1{:}n_{\mathrm{sh}}\}$ **do**
        $x_1 := \sum_{k_1,k_2\in\{1:d\}} \mathtt{dtheta\_dK}(k_1,nj,l,m) * \mathtt{dif\_l}(k_1,k_2) * \mathtt{dtheta\_dK}(k_2,ni,l,m)$
        $x_2 := \mathtt{theta}(ni,l) * \sum_{k_1\in\{1:d\}} \mathtt{beta\_l}(k_1) * \mathtt{dtheta\_dK}(k_1,nj,l,m)$
        $x_3 := \mathtt{theta}(ni,l) * \mathtt{mu\_l} * \mathtt{theta}(nj,l)$
        $\mathtt{tmp}(ni,nj) := \mathtt{tmp}(ni,nj) + [x_1+x_2+x_3] * \mathtt{weight\_K}(l,m)$
      end for
    end for
  end for
  Accumulate tmp in $\mathcal{A}_{\mathcal{Q}}$ as in Algorithm 29.2
end for

Algorithm 30.2 Assembling of RHS vector $\mathsf{B}_{\mathcal{Q}}$.

$\mathsf{B}_{\mathcal{Q}} := 0$
for $m \in \{1{:}N_{\mathrm{c}}\}$ **do**
  for $l \in \{1{:}l_{\mathcal{Q}}\}$ **do**; $\mathtt{tmp} := 0$
    for $k_1 \in \{1{:}d\}$ **do**
      $\mathtt{xi\_l}(k_1) := \sum_{n\in\{1:n_{\mathrm{geo}}\}} \mathtt{coord}(k_1,\mathtt{j\_geo}(n,m))\,\mathtt{psi}(n,l)$
    end for
    for $ni \in \{1{:}n_{\mathrm{sh}}\}$ **do**
      $x_1 := f(\mathtt{xi\_l}) * \mathtt{theta}(ni,l)$
      $x_2 := \sum_{k_1\in\{1:d\}} \boldsymbol{\beta}_{k_1}(\mathtt{xi\_l}) * \mathtt{dtheta\_dK}(k_1,ni,l,m)$
      $\mathtt{tmp}(ni) := \mathtt{tmp}(ni) + [x_1+x_2] * \mathtt{weight\_K}(l,m)$
    end for
  end for
  for $ni \in \{1{:}n_{\mathrm{sh}}\}$ **do**; $i := \mathtt{j\_dof}(m,ni)$
    $\mathsf{B}_{\mathcal{Q},i} := \mathsf{B}_{\mathcal{Q},i} + \mathtt{tmp}(ni)$
  end for
end for

Chapter 31

Scalar second-order elliptic PDEs

Exercises

Exercise 31.1 (Cordes). Prove that ellipticity implies the Cordes condition if $d = 2$. (*Hint*: use that $\|\mathbb{d}\|_F^2 = (\mathrm{tr}(\mathbb{d}))^2 - 2\det(\mathbb{d})$.)

Exercise 31.2 (Poincaré–Steklov). Prove (31.23). (*Hint*: use (3.12).)

Exercise 31.3 (Potential flow). Consider the PDE $\nabla\cdot(-\kappa\nabla u+\boldsymbol{\beta}u) = f$ in D with homogeneous Dirichlet conditions and assume that κ is a positive real number. Assume that $\boldsymbol{\beta} := \nabla\psi$ for some smooth function ψ (we say that $\boldsymbol{\beta}$ is a potential flow). Find a functional $\mathfrak{E} : H_0^1(D) \to \mathbb{R}$ of which the weak solution u is a minimizer on $H_0^1(D)$. (*Hint*: consider the function $e^{-\psi/\kappa}u$.)

Exercise 31.4 (Purely diffusive Neumann). Prove Proposition 31.19. (*Hint*: for all $w \in H^1(D)$, the function $\tilde{w} := w - \underline{w}_D$ is in $H_*^1(D)$, use also the Poincaré–Steklov inequality from Lemma 3.24.)

Exercise 31.5 (Mixed Dirichlet–Neumann). The goal is to show by a counterexample that one cannot assert that the weak solution is in $H^2(D)$ for the mixed Dirichlet–Neumann problem even if the domain and the boundary data are smooth. Using polar coordinates, set $D := \{(r,\theta) \in (0,1)\times(0,\pi)\}$, $\partial D_{\mathrm{n}} := \{r \in (0,1),\ \theta = \pi\}$, and $\partial D_{\mathrm{d}} := \partial D\backslash\partial D_{\mathrm{n}}$. Verify that the function $u(r,\theta) := r^{\frac{1}{2}}\sin(\frac{1}{2}\theta)$ satisfies $-\Delta u = 0$ in D, $\frac{\partial u}{\partial n}|_{D_{\mathrm{n}}} = 0$, and $u_{|D_{\mathrm{d}}} = r^{\frac{1}{2}}\sin(\frac{1}{2}\theta)$. (*Hint*: in polar coordinates, $\Delta u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial\theta^2}$.) Verify that $u \notin H^2(D)$.

Exercise 31.6 ($H^2(\mathbb{R}^d)$-seminorm). Prove that $|\phi|_{H^2(\mathbb{R}^d)} = \|\Delta\phi\|_{L^2(\mathbb{R}^d)}$ for all $\phi \in C_0^\infty(\mathbb{R}^d)$. (*Hint*: use Theorem B.3.)

Exercise 31.7 (Counterexample to elliptic regularity in $W^{2,\infty}(D)$). Let D be the unit disk in $\mathbb{R}^2$. Consider the function $u(x_1,x_2) := x_1x_2\ln(r)$ with $r^2 := x_1^2 + x_2^2$ (note that $u_{|\partial D} = 0$). Verify that $\Delta u \in L^\infty(D)$, but that $u \notin W^{2,\infty}(D)$. (*Hint*: consider the cross-derivative.)

Exercise 31.8 (Domain with slit). Let $D := \{r \in (0,1),\ \theta \in (0,2\pi)\}$, where (r,θ) are the polar coordinates, i.e., $\overline{D}$ is the closed ball of radius 1 centered at 0. Let $u(r,\theta) := r\cos(\frac{1}{2}\theta)$ for all $r > 0$ and $\theta \in [0,2\pi)$. (i) Let $p \in [1,\infty)$. Is $u_{|D}$ in $W^{1,p}(D)$? Is $u_{|\mathrm{int}(\overline{D})}$ in $W^{1,p}(\mathrm{int}(\overline{D}))$? (*Hint*: recall Example 4.3.) (ii) Is the restriction to D of the functions in $C^1(\overline{D})$ dense in $W^{1,p}(D)$? (*Hint*: argue by contradiction and use that $\|v_{|D}\|_{W^{1,p}(D)} = \|v_{|\mathrm{int}(\overline{D})}\|_{W^{1,p}(\mathrm{int}(\overline{D}))}$ for all $v \in C^1(\overline{D})$.)

Exercise 31.9 (A priori estimate). Consider the PDE $-\kappa_0\Delta u + \boldsymbol{\beta}\cdot\nabla u + \mu_0 u = f$ with homogeneous Dirichlet conditions. Assume that $\kappa_0, \mu_0 \in \mathbb{R}$, $\kappa_0 > 0$, $\nabla\cdot\boldsymbol{\beta} = 0$, $\boldsymbol{\beta}_{|\partial D} = \mathbf{0}$, and $f \in H^1_0(D)$. Let $\nabla_{\mathrm{s}}\boldsymbol{\beta} := \frac{1}{2}(\nabla\boldsymbol{\beta} + (\nabla\boldsymbol{\beta})^{\mathsf{T}})$ denote the symmetric part of the gradient of $\boldsymbol{\beta}$, and assume that there is $\mu_0' > 0$ s.t. $\nabla_{\mathrm{s}}\boldsymbol{\beta} + \mu_0\mathbb{I}_d \geq \mu_0'\mathbb{I}_d$ in the sense of quadratic forms. Prove that $|u|_{H^1(D)} \leq (\mu_0')^{-1}|f|_{H^1(D)}$ and $\|\Delta u\|_{L^2(D)} \leq (4\mu_0'\kappa_0)^{-\frac{1}{2}}|f|_{H^1(D)}$. (*Hint*: use $-\Delta u$ as a test function.) *Note*: these results are established in Beirão da Veiga [3], Burman [8].

Exercise 31.10 (Complex-valued diffusion). Assume that the domain D is partitioned into two disjoint subdomains D_1 and D_2. Let κ_1, κ_2 be two complex numbers, both with positive modulus and such that $\frac{\kappa_1}{\kappa_2} \notin \mathbb{R}_-$. Set $\kappa(x) := \kappa_1\mathbb{1}_{D_1}(\boldsymbol{x}) + \kappa_2\mathbb{1}_{D_2}(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$. Let $f \in L^2(D)$. Show that the problem of seeking $u \in V := H^1_0(D;\mathbb{C})$ such that $a(u,w) := \int_D \kappa\nabla u\cdot\nabla\overline{w}\,\mathrm{d}x = \int_D f\overline{w}\,\mathrm{d}x$ for all $w \in V$ is well-posed. (*Hint*: use (25.7).)

Exercise 31.11 (Dependence on diffusion coefficient). Consider two numbers $0 < \lambda_\flat \leq \lambda_\sharp < \infty$ and define the set $K := \{\kappa \in L^\infty(D;\mathbb{R}) \mid \kappa(\boldsymbol{x}) \in [\lambda_\flat, \lambda_\sharp], \text{a.e. } \boldsymbol{x} \in D\}$. Let $V := H^1_0(D)$ equipped with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$ and $V' = H^{-1}(D)$. Consider the operator $T_\kappa : V \to V'$ s.t. $T_\kappa(v) := -\nabla\cdot(\kappa\nabla v)$ for all $v \in V$ and all $\kappa \in K$. (i) Prove that $\lambda_\flat \leq \|T_\kappa\|_{\mathcal{L}(V;V')} \leq \lambda_\sharp$ and that T_κ is an isomorphism. (*Hint*: use Proposition 31.8 with $\theta := 1$ and the bilinear form $a(v,w) := \int_D \kappa\nabla v\cdot\nabla w\,\mathrm{d}x$ on $V \times V$.) (ii) Prove that $\|T_\kappa - T_{\kappa'}\|_{\mathcal{L}(V;V')} = \|\kappa - \kappa'\|_{L^\infty(D)}$ for all $\kappa, \kappa' \in K \cap C^0(D;\mathbb{R})$. (*Hint*: if $\|\kappa-\kappa'\|_{L^\infty(D)} > 0$, for all $\epsilon > 0$ there is an open subset $D_\epsilon \subset D$ such that the sign of $(\kappa-\kappa')_{|D_\epsilon}$ is constant and $|\kappa-\kappa'| \geq \|\kappa-\kappa'\|_{L^\infty(D)} - \epsilon$ in D_ϵ; then consider functions in $H^1_0(D_\epsilon)$.) (iii) Let $S_\kappa := T_\kappa^{-1} \in \mathcal{L}(V';V)$. Prove that $\lambda_\flat^2\|S_\kappa - S_{\kappa'}\|_{\mathcal{L}(V';V)} \leq \|\kappa - \kappa'\|_{L^\infty(D)} \leq \lambda_\sharp^2\|S_\kappa - S_{\kappa'}\|_{\mathcal{L}(V';V)}$ for all $\kappa, \kappa' \in K \cap C^0(D;\mathbb{R})$. (*Hint*: $S_\kappa - S_{\kappa'} = S_\kappa(T_{\kappa'} - T_\kappa)S_{\kappa'}$.)

Solution to exercises

Exercise 31.1 (Cordes). Using the symmetry of $\mathbb{d}$, we have $\|\mathbb{d}\|_F^2 = (\mathrm{tr}(\mathbb{d}))^2 - 2\det(\mathbb{d})$, where $\det(\mathbb{d})$ is the determinant of $\mathbb{d}$, so that $\frac{\|\mathbb{d}\|_F^2}{(\mathrm{tr}(\mathbb{d}))^2} = \frac{1}{1+\epsilon}$ with $\epsilon := \frac{2\det(\mathbb{d})}{\|\mathbb{d}\|_F^2}$. Since $\det(\mathbb{d}) > 0$ by the ellipticity condition, we have $\epsilon > 0$. Since $(\mathrm{tr}(\mathbb{d}))^2 \geq 4\det(\mathbb{d})$ if $d = 2$, we have $\|\mathbb{d}\|_F^2 \geq 2\det(\mathbb{d})$, so that $\epsilon \leq 1$, the case $\epsilon = 1$ being reached when both eigenvalues of $\mathbb{d}$ are equal, i.e., $\mathbb{d} = \lambda\mathbb{I}$ with $\lambda > 0$.

Exercise 31.2 (Poincaré–Steklov). Let us define the linear form $f(v) := \ell_D^{\frac{1}{2}}|\partial D|^{-\frac{1}{2}}\int_{\partial D}\gamma^{\mathrm{g}}(v)\,\mathrm{d}s$. This defines a bounded linear form on $H^1(D)$. Applying (3.12) (with $p := 2$), we infer that there is $\check{C}_{\mathrm{PS}}$ s.t.

$$\sqrt{2}\check{C}_{\mathrm{PS}}\|v\|_{L^2(D)} \leq \ell_D\|\nabla v\|_{\boldsymbol{L}^2(D)} + \ell_D^{\frac{1}{2}}|\partial D|^{-\frac{1}{2}}\left|\int_{\partial D}\gamma^{\mathrm{g}}(v)\,\mathrm{d}s\right|,$$

for all $v \in H^1(D)$. The rightmost term is bounded as $|\int_{\partial D}\gamma^{\mathrm{g}}(v)\,\mathrm{d}s| \leq |\partial D|^{\frac{1}{2}}\|\gamma^{\mathrm{g}}(v)\|_{L^2(\partial D)}$ owing to the Cauchy–Schwarz inequality. Hence, we have

$$\sqrt{2}\check{C}_{\mathrm{PS}}\|v\|_{L^2(D)} \leq \ell_D(\|\nabla v\|_{\boldsymbol{L}^2(D)} + \ell_D^{-\frac{1}{2}}\|\gamma^{\mathrm{g}}(v)\|_{L^2(\partial D)}),$$

and we conclude using Young's inequality: $(a+b) \leq (2(a^2+b^2))^{\frac{1}{2}}$.

Exercise 31.3 (Potential flow). We observe that

$$\nabla\left(e^{-\psi/\kappa}u\right) = \frac{1}{\kappa}e^{-\psi/\kappa}(\kappa\nabla u - \boldsymbol{\beta}u),$$

since $\nabla\psi = \boldsymbol{\beta}$. Consider the functional $\mathfrak{E} : H^1_0(D) \to \mathbb{R}$ such that

$$\mathfrak{E}(v) := \frac{1}{2}\int_D e^{\psi/\kappa}\kappa \left|\nabla\left(e^{-\psi/\kappa}v\right)\right|^2 \mathrm{d}x - \int_D e^{-\psi/\kappa} f v \,\mathrm{d}x,$$

for all $v \in H^1_0(D)$. Proceeding as in the proof of Proposition 25.8, we see that $u \in H^1_0(D)$ is a global minimizer of $\mathfrak{E}$ if and only if

$$\int_D e^{\psi/\kappa}\kappa\nabla\left(e^{-\psi/\kappa}u\right)\cdot\nabla\left(e^{-\psi/\kappa}w\right)\mathrm{d}x = \int_D e^{-\psi/\kappa} f w \,\mathrm{d}x,$$

for all $w \in H^1_0(D)$. Taking w to be arbitrary in $C^\infty_0(D)$, we infer that

$$\int_D e^{-\psi/\kappa}\nabla\cdot(-\kappa\nabla u + \boldsymbol{\beta}u)w \,\mathrm{d}x = \int_D e^{-\psi/\kappa} f w \,\mathrm{d}x,$$

which shows that u satisfies the PDE $\nabla\cdot(-\kappa\nabla u + \boldsymbol{\beta}u) = f$ a.e. in D.

Exercise 31.4 (Purely diffusive Neumann). For all $w \in H^1(D)$, writing $w := \tilde{w} + \underline{w}_D$ with $\tilde{w} \in H^1_*(D)$ and testing the weak formulation against $\tilde{w}$, we infer that the weak solution satisfies

$$\begin{aligned} a_{\mathrm{d}}(u,w) = a_{\mathrm{d}}(u,\tilde{w}) &= \int_D f\tilde{w}\,\mathrm{d}x + \int_{\partial D} g\gamma^{\mathrm{g}}(\tilde{w})\,\mathrm{d}s \\ &= \int_D f w\,\mathrm{d}x + \int_{\partial D} g\gamma^{\mathrm{g}}(w)\,\mathrm{d}s, \end{aligned}$$

where we used the compatibility condition (31.28) in the last equality. Since the equality $a_{\mathrm{d}}(u,w) = \int_D f w\,\mathrm{d}x + \int_{\partial D} g\gamma^{\mathrm{g}}(w)\,\mathrm{d}s$ is valid for every function $w \in H^1(D)$, we infer as in the case of Robin conditions that the PDE and the boundary condition in (31.27) are satisfied a.e. in D and a.e. on ∂D, respectively. To prove the well-posedness of the weak formulation, we use the Poincaré–Steklov Lemma 3.24 with $p := 2$, i.e., $C_{\mathrm{PS}}\|v\|_{L^2(D)} \le \ell_D\|\nabla v\|_{\boldsymbol{L}^2(D)}$ for all $v \in H^1_*(D)$, so that $V := H^1_*(D)$ equipped with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$ is a Hilbert space. Since $a_{\mathrm{d}}(v,v) \ge \lambda_\flat\|v\|_V^2$, this proves the coercivity of a_{d}. Finally, the well-posedness follows from the Lax–Milgram lemma.

Exercise 31.5 (Mixed Dirichlet–Neumann). A direct computation gives

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) = \frac{1}{4}r^{-\frac{3}{2}}\sin\left(\frac{1}{2}\theta\right), \qquad \frac{1}{r^2}\frac{\partial^2 u}{\partial\theta^2} = -\frac{1}{4}r^{-\frac{3}{2}}\sin\left(\frac{1}{2}\theta\right),$$

so that $\Delta u = 0$. The Dirichlet condition is clearly satisfied on ∂D_{d}. Concerning the Neumann condition on ∂D_{n}, we observe that $\frac{\partial u}{\partial n} = \frac{1}{r}\frac{\partial u}{\partial\theta} = \frac{1}{2}r^{-\frac{1}{2}}\cos(\frac{1}{2}\theta)$ which vanishes for $\theta = \pi$. Finally, we observe that $\frac{\partial^2 u}{\partial r^2} = -\frac{1}{4}r^{-\frac{3}{2}}\sin(\frac{1}{2}\theta)$ and that $\int_0^1 r^{-3} r\,\mathrm{d}r$ is not bounded.

Exercise 31.6 ($H^2(\mathbb{R}^d)$-seminorm). Let $\phi \in C^\infty_0(\mathbb{R}^d)$. Integrating by parts, we infer that

$$\begin{aligned} |\phi|^2_{H^2(\mathbb{R}^d)} &= \sum_{i,j\in\{1:d\}}\int_{\mathbb{R}^d}\frac{\partial^2\phi}{\partial x_i\partial x_j}\frac{\partial^2\phi}{\partial x_i\partial x_j}\,\mathrm{d}x \\ &= -\sum_{i,j\in\{1:d\}}\int_{\mathbb{R}^d}\frac{\partial\phi}{\partial x_i}\frac{\partial}{\partial x_j}\left(\frac{\partial^2\phi}{\partial x_i\partial x_j}\right)\mathrm{d}x \\ &= -\sum_{i,j\in\{1:d\}}\int_{\mathbb{R}^d}\frac{\partial\phi}{\partial x_i}\frac{\partial^3\phi}{\partial x_i\partial x_j^2}\,\mathrm{d}x \\ &= \sum_{i,j\in\{1:d\}}\int_{\mathbb{R}^d}\frac{\partial^2\phi}{\partial x_i^2}\frac{\partial^2\phi}{\partial x_j^2}\,\mathrm{d}x = \|\Delta\phi\|^2_{L^2(\mathbb{R}^d)}, \end{aligned}$$

where we used Theorem B.3 to exchange the order of the partial derivatives.

Exercise 31.7 (Counterexample to elliptic regularity in $W^{2,\infty}(D)$). We observe that

$$\begin{aligned}\Delta u &= x_1x_2\Delta(\ln(r)) + 2\nabla(x_1x_2)\cdot\nabla(\ln(r)) + \Delta(x_1x_2)\ln(r)\\ &= 2\nabla(x_1x_2)\cdot\nabla(\ln(r)) = 4\frac{x_1x_2}{r^2},\end{aligned}$$

so that $\Delta u \in L^\infty(D)$. Moreover, we have

$$\frac{\partial^2 u}{\partial x_1\partial x_2} = \ln(r) + 1 - \frac{2x_1^2x_2^2}{r^4},$$

which is unbounded at the origin.

Exercise 31.8 (Domain with slit). (i) Since $\partial_\theta u = -\frac{1}{2}r\sin(\frac{1}{2}\theta) + 2r\delta_{\theta=0}$, where $\delta_{\theta=0}$ is the Dirac measure whose support is the segment $\{r\in(0,1),\ \theta=0\} = \{x_1\in(0,1),\ x_2=0\}$, we infer that $u_{|D}\in W^{1,p}(D)$, but $u_{|\mathrm{int}(\overline{D})}\notin W^{1,p}(\mathrm{int}(\overline{D}))$ since $\delta_{\theta=0}$ cannot be identified with any function in $L^p(\mathrm{int}(\overline{D}))$; see Example 4.3.
(ii) Assume that the restriction to D of the functions in $C^1(\overline{D})$ is dense in $W^{1,p}(D)$. Since $u_{|D}\in W^{1,p}(D)$, there is a sequence of functions in $C^1(\overline{D})$, say $(v_n)_{n\in\mathbb{N}}$, such that $v_{n|D}\to u_{|D}$ in $W^{1,p}(D)$. But, since $v_n\in C^1(\overline{D})\subset W^{1,p}(\overline{D})$ and $|\mathrm{int}(\overline{D})\backslash D| = 0$, we have

$$\|v_{n|D}\|_{W^{1,p}(D)} = \|v_{n|\mathrm{int}(\overline{D})}\|_{W^{1,p}(\mathrm{int}(\overline{D}))}.$$

This means that $(v_{n|\mathrm{int}(\overline{D})})_{n\in\mathbb{N}}$ is a Cauchy sequence in $W^{1,p}(\mathrm{int}(\overline{D}))$. Let w be the limit in question. We have

$$w_{|D} = u_{|D},\quad \text{a.e. in } D.$$

This proves that $w_{|\mathrm{int}(\overline{D})} = u_{|\mathrm{int}(\overline{D})}$ since $|\,\mathrm{int}(\overline{D})\backslash D| = 0$. This, in turn, establishes that $u_{|\mathrm{int}(\overline{D})}\in W^{1,p}(\mathrm{int}(\overline{D}))$, which is a contradiction. Hence, the restriction to D of the functions in $C^1(\overline{D})$ is not dense in $W^{1,p}(D)$.

Exercise 31.9 (A priori estimate). Following the hint and integrating by parts, we infer that

$$\kappa_0\|\Delta u\|_{L^2(D)}^2 - (\boldsymbol{\beta}\cdot\nabla u,\Delta u)_{L^2(D)} + \mu_0|u|_{H^1(D)}^2 = -(f,\Delta u)_{L^2(D)} = (\nabla f,\nabla u)_{\boldsymbol{L}^2(D)},$$

where we used that $u\in H_0^1(D)$ in the third term on the left-hand side and $f\in H_0^1(D)$ on the right-hand side. Using that $\boldsymbol{\beta}_{|\partial D} = \mathbf{0}$, we infer that

$$\begin{aligned}-(\boldsymbol{\beta}\cdot\nabla u,\Delta u)_{L^2(D)} &= -\sum_{i,j\in\{1:d\}}(\beta_i\partial_i u,\partial_j\partial_j u)_{L^2(D)}\\ &= \sum_{i,j\in\{1:d\}}((\partial_j\beta_i)\partial_i u,\partial_j u)_{L^2(D)} + (\beta_i\partial_i(\partial_j u),\partial_j u)_{L^2(D)}\\ &=: \mathfrak{T}_1+\mathfrak{T}_2.\end{aligned}$$

We have $\mathfrak{T}_1 = ((\nabla_{\mathrm{s}}\boldsymbol{\beta})\nabla u,\nabla u)_{\boldsymbol{L}^2(D)}$. Using that $\nabla\cdot\boldsymbol{\beta} = 0$ and using again that $\boldsymbol{\beta}$ vanishes at the boundary, we obtain that

$$\mathfrak{T}_2 = \sum_{i,j\in\{1:d\}}(\boldsymbol{\beta}\cdot\nabla\partial_j u,\partial_j u)_{L^2(D)} = \int_D\frac{1}{2}\nabla\cdot(\boldsymbol{\beta}\|\nabla u\|^2)\,\mathrm{d}x = 0.$$

In summary, we have shown that

$$\kappa_0\|\Delta u\|_{L^2(D)}^2 + ((\nabla_{\mathrm{s}}\boldsymbol{\beta})\nabla u,\nabla u)_{\boldsymbol{L}^2(D)} + \mu_0|u|_{H^1(D)}^2 = (\nabla f,\nabla u)_{\boldsymbol{L}^2(D)}.$$

Our assumption on $\nabla_{\mathrm{s}}\boldsymbol{\beta}$ implies that

$$\kappa_0\|\Delta u\|_{L^2(D)}^2+\mu_0'|u|_{H^1(D)}^2\le(\nabla f,\nabla u)_{\boldsymbol{L}^2(D)}.$$

The estimate on $|u|_{H^1(D)}$ follows by applying the Cauchy–Schwarz inequality to the right-hand side. The estimate on $\|\Delta u\|_{L^2(D)}$ follows by bounding the right-hand side as $\mu_0'|u|_{H^1(D)}^2+(4\mu_0')^{-1}|f|_{H^1(D)}^2$.

Exercise 31.10 (Complex-valued diffusion). Let us write $\kappa_m:=|\kappa_m|e^{\mathrm{i}\varphi_m}$ for all $m\in\{1,2\}$. Set $\xi:=e^{-\mathrm{i}\frac{\varphi_1+\varphi_2}{2}}$. Then the real part of $\xi\kappa_1$ is $|\kappa_1|\cos\left(\frac{\varphi_1-\varphi_2}{2}\right)$ and that of $\xi\kappa_2$ is $|\kappa_2|\cos\left(\frac{\varphi_2-\varphi_1}{2}\right)$. It is readily seen that these two real numbers have the same sign and that they are both nonzero since $\frac{\varphi_2-\varphi_1}{2}\neq\pm\frac{\pi}{2}$ (since otherwise $\frac{\kappa_1}{\kappa_2}$ would be a negative real number). Hence, up to a possible sign change in ξ, the bilinear form a satisfies the coercivity property (25.7). We conclude by invoking the Lax–Milgram lemma.

Exercise 31.11 (Dependence on diffusion coefficient). (i) We have

$$\|T_\kappa(v)\|_{V'}=\sup_{w\in V}\frac{|\langle\nabla\cdot(\kappa\nabla v),w\rangle_{V',V}|}{\|w\|_V}=\sup_{w\in V}\frac{|\int_D\kappa\nabla v\cdot\nabla w\,\mathrm{d}x|}{\|w\|_V},$$

for all $v\in V$. Recalling the definition of the $\|\cdot\|_V$-norm, this implies that $\|T_\kappa(v)\|_{V'}\le\lambda_\sharp\|v\|_V$ and that

$$\|T_\kappa(v)\|_{V'}\ge\frac{|\int_D\kappa\nabla v\cdot\nabla v\,\mathrm{d}x|}{\|v\|_V}\ge\lambda_\flat\|v\|_V.$$

This lower bound proves that T_κ is injective, and the above two bounds together prove that

$$\lambda_\flat\le\sup_{v\in V}\frac{\|T_\kappa(v)\|_{V'}}{\|v\|_V}=\|T_\kappa\|_{\mathcal{L}(V;V')}\le\lambda_\sharp.$$

It remains to prove that T_κ is surjective. Proposition 31.8 applied with $\theta:=1$ (and $\mu:=0$, $\boldsymbol{\beta}:=\boldsymbol{0}$) implies that the bilinear form $a(v,w):=\int_D\kappa\nabla v\cdot\nabla w\,\mathrm{d}x$ is coercive on V with $a(v,v)\ge\lambda_\flat\|v\|_V^2$. The Lax–Milgram lemma then implies that for all $\phi\in V'$, there is a unique $v_\phi\in V$ s.t. $a(v_\phi,w)=\langle\phi,w\rangle_{V',V}$ for all $w\in V$. Let $\phi\in V'$. Then we have for all $w\in V$,

$$\langle T_\kappa(v_\phi),w\rangle_{V',V}=-\langle\nabla\cdot(\kappa\nabla v_\phi),w\rangle_{V',V}=\int_D\kappa\nabla v_\phi\cdot\nabla w\,\mathrm{d}x=a(v_\phi,w)=\langle\phi,w\rangle_{V',V}.$$

This shows that $T_\kappa(v_\phi)=\phi$, i.e., T_κ is surjective.
(ii) We have

$$\|T_\kappa-T_{\kappa'}\|_{\mathcal{L}(V;V')}=\sup_{v\in V}\sup_{w\in V}\frac{|\int_D(\kappa-\kappa')\nabla v\cdot\nabla w\,\mathrm{d}x|}{\|v\|_V\|w\|_V}\le\|\kappa-\kappa'\|_{L^\infty(D)}.$$

Assume that $\|\kappa-\kappa'\|_{L^\infty(D)}>0$ since otherwise there is nothing to prove. Let $\epsilon>0$ and assume that $\epsilon\le\|\kappa-\kappa'\|_{L^\infty(D)}$. There is a measurable subset $E_\epsilon\subset D$ with $|E_\epsilon|>0$ s.t. the sign of $(\kappa-\kappa')_{|E_\epsilon}$ is constant in E_ϵ and $|\kappa-\kappa'|\ge\|\kappa-\kappa'\|_{L^\infty(D)}-\epsilon$ in E_ϵ. Since we are assuming that κ,κ' are continuous functions, there is an open subset $D_\epsilon\subset E_\epsilon$. Observing that the zero-extension of a function in $H_0^1(D_\epsilon)$ is in V, we infer that

$$\|T_\kappa-T_{\kappa'}\|_{\mathcal{L}(V;V')}\ge\sup_{v\in H_0^1(D_\epsilon)}\frac{|\int_{D_\epsilon}(\kappa-\kappa')\|\nabla v\|_{\ell^2}^2\,\mathrm{d}x|}{\int_{D_\epsilon}\|\nabla v\|_{\ell^2}^2\,\mathrm{d}x}\ge\|\kappa-\kappa'\|_{L^\infty(D)}-\epsilon.$$

Since $\epsilon > 0$ is arbitrary, this proves that $\|T_\kappa - T_{\kappa'}\|_{\mathcal{L}(V;V')} = \|\kappa - \kappa'\|_{L^\infty(D)}$.
(iii) Since $S_\kappa = T_\kappa^{-1}$, we infer from the bounds derived in Step (i) that

$$\lambda_\sharp^{-1} \le \|S_\kappa\|_{\mathcal{L}(V';V)} \le \lambda_\flat^{-1}.$$

Using the hint, we obtain

$$\begin{aligned}\|S_\kappa - S_{\kappa'}\|_{\mathcal{L}(V';V)} &\le \|S_\kappa\|_{\mathcal{L}(V';V)}\|T_\kappa - T_{\kappa'}\|_{\mathcal{L}(V;V')}\|S_{\kappa'}\|_{\mathcal{L}(V';V)} \\ &\le \lambda_\flat^{-2}\|\kappa - \kappa'\|_{L^\infty(D)}.\end{aligned}$$

This proves that $\lambda_\flat^2\|S_\kappa - S_{\kappa'}\|_{\mathcal{L}(V';V)} \le \|\kappa - \kappa'\|_{L^\infty(D)}$. Finally, using the identity $T_\kappa - T_{\kappa'} = T_\kappa(S_{\kappa'} - S_\kappa)T_{\kappa'}$ and reasoning similarly proves that $\|\kappa - \kappa'\|_{L^\infty(D)} \le \lambda_\sharp^2\|S_\kappa - S_{\kappa'}\|_{\mathcal{L}(V';V)}$.

Chapter 32

H^1-conforming approximation (I)

Exercises

Exercise 32.1 (Discrete solution map). Let G_h be defined in (32.6). (i) Prove that $\|\nabla(v - G_h(v))\|_{\boldsymbol{L}^2(D)} \le ch^r|v|_{H^{1+r}(D)}$ for all $r \in (0,k]$, all $v \in H^{1+r}(D)$, and all $h \in \mathcal{H}$. (*Hint*: observe that $G_h(\mathcal{I}_{h0}^{\mathrm{g,av}}(v)) = \mathcal{I}_{h0}^{\mathrm{g,av}}(v)$.) (ii) Assume that the adjoint operator A^* has a smoothing property in $H^{1+s}(D)$ for some real number $s \in (0,1]$. Prove that $\|v - G_h(v)\|_{L^2(D)} \le ch^{r+s}\ell_D^{1-s}|v|_{H^{1+r}(D)}$. (*Hint*: consider the adjoint problem $A^*(\zeta) = v - G_h(v)$.)

Exercise 32.2 (H^{-1}-estimate). Assume that for all $g \in H^1(D)$, the adjoint solution $\zeta \in H_0^1(D)$ s.t. $A^*(\zeta) = g$ satisfies $\|\zeta\|_{H^{2+s}(D)} \le c_{\mathrm{smo}}\alpha^{-1}\ell_D^2\|g\|_{H^1(D)}$ with $s \in (\frac{1}{2}, 1]$. Assume that $k \ge 1 + s$. Let $\|v\|_{H^{-1}(D)} := \sup_{z\in H_0^1(D)} \frac{(v,z)_{L^2(D)}}{|z|_{H^1(D)}}$ for all $v \in L^2(D)$. Prove that $\|u - u_h\|_{H^{-1}(D)} \le ch^{1+s}\ell_D^{1-s}\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)}$. (*Hint*: consider the adjoint problem $A^*(\zeta) = z$.)

Exercise 32.3 (Compactness). The goal is to prove Theorem 32.8. Let $I : V \to L$ be the natural embedding and define $\epsilon(h) := \sup_{v\in V\setminus V_h} \frac{\|G_h(v)-v\|_L}{\|G_hv-v\|_V}$. (i) Prove that $\|G_h - I\|_{\mathcal{L}(V;L)} \le \frac{\|a\|}{\alpha}\epsilon(h)$, where α and $\|a\|$ are the coercivity and the boundedness constants of a on $V \times V$. (ii) Assume that $\lim_{h\to 0}\epsilon(h) = 0$. Prove that I is compact. (*Hint*: use (i).) (iii) Let $R : L \to V$ be s.t. $a(y, R(f)) := (y, f)_L$ for all $y \in V$ and all $f \in L$. Assuming that I is compact, prove that R is compact. (*Hint*: prove that $R = (A^*)^{-1}I^*$ and use Schauder's theorem; see Theorem C.48.) (iv) Let $P_h^V : V \to V_h$ be the V-orthogonal projection onto V_h. Let $R_h : L \to V_h$ be the operator defined by $a(v_h, R_h(f)) := (v_h, f)_L$, for all $v_h \in V_h$ and all $f \in L$. Prove that $\|R - R_h\|_{\mathcal{L}(L;V)} \le \frac{\|a\|}{\alpha}\|R - P_h^V \circ R\|_{\mathcal{L}(L;V)}$. (v) Assuming that I is compact, prove that $\lim_{h\to 0}\|R - R_h\|_{\mathcal{L}(L;V)} = 0$. (*Hint*: use (iii)-(iv) and proceed as in Remark C.5.) (vi) Assuming that I is compact, prove that $\lim_{h\to 0}\epsilon(h) = 0$.

Exercise 32.4 (Source approximation). Let $f \in L^2(D)$, let $\mathcal{I}_h^{\mathrm{b}}(f)$ be the L^2-projection of f onto $P_{k'}^{\mathrm{b}}(\mathcal{T}_h)$. Consider the discrete problem (32.5) with the right-hand side $\int_D \mathcal{I}_h^{\mathrm{b}}(f)w_h\,\mathrm{d}x$, that is: Find $u_h \in V_h := P_{k,0}^{\mathrm{g}}(\mathcal{T}_h)$ s.t. $a(u_h, w_h) = \ell_h(w_h) := \int_D \mathcal{I}_h^{\mathrm{b}}(f)w_h\,\mathrm{d}x$ for all $w_h \in V_h$. (i) How should (32.7) be rewritten? Show that $k' := k - 1$ leads to an optimal H^1-norm error estimate. (ii) How should (32.19) be rewritten? Assuming full elliptic regularity, show that $k' := k$ leads to an optimal L^2-norm error estimate.

Exercise 32.5 (Advection-diffusion, 1D). Let $D := (0,1)$. Let ν, b be positive real numbers. Let $f : D \to \mathbb{R}$ be a smooth function. Consider the model problem $-\nu u'' + bu' = f$ in D, $u(0) = 0$, $u(1) = 0$. Consider H^1-conforming $\mathbb{P}_1$ Lagrange finite elements on the uniform grid $\mathcal{T}_h$ with nodes $x_i := ih$, $\forall i \in \{0{:}I\}$, and meshsize $h := \frac{1}{I+1}$. (i) Evaluate the stiffness matrix. (*Hint*: factor out the ratio $\frac{\nu}{h}$ and introduce the local Péclet number $\gamma := \frac{bh}{\nu}$.) (ii) Solve the linear system when $f := 1$ and plot the solutions for $h := 10^{-2}$ and $\gamma \in \{0.1, 1, 10\}$. (*Hint*: wrtite $\mathsf{U} = \mathsf{U}^0 + \tilde{\mathsf{U}} \in \mathbb{R}^I$ with $\mathsf{U}^0_i := b^{-1}ih$ and $\tilde{\mathsf{U}}_i := \varrho + \theta\delta^i$ for some constants ϱ, θ, δ.) (iii) Consider now the boundary conditions $u(0) = 0$ and $u'(1) = 0$. Write the weak formulation and show its well-posedness. Evaluate the stiffness matrix. (*Hint*: the matrix is of order $(I+1)$.) Derive the equation satisfied by $h^{-1}(\mathsf{U}_{I+1} - \mathsf{U}_I)$, and find the limit values as $h \to 0$ with fixed $\nu > 0$ and as $\nu \to 0$ with fixed $h \in \mathcal{H}$.

Solution to exercises

Exercise 32.1 (Discrete solution map). (i) We observe that

$$\begin{aligned}\|\nabla(v - G_h(v))\|_{\boldsymbol{L}^2(D)} &\le \|\nabla(v - \mathcal{I}^{\mathrm{g,av}}_{h0}(v))\|_{\boldsymbol{L}^2(D)} + \|\nabla G_h(v - \mathcal{I}^{\mathrm{g,av}}_{h0}(v))\|_{\boldsymbol{L}^2(D)}\\ &\le c\,\|\nabla(v - \mathcal{I}^{\mathrm{g,av}}_{h0}(v))\|_{\boldsymbol{L}^2(D)} \le c\,h^r|v|_{H^{1+r}(D)},\end{aligned}$$

where we used the triangle inequality, the fact that $G_h(\mathcal{I}^{\mathrm{g,av}}_{h0}(v)) = \mathcal{I}^{\mathrm{g,av}}_{h0}(v)$, the H^1-stability of G_h, and Corollary 22.16.
(ii) Consider the adjoint problem which consists of seeking $\zeta \in H^1_0(D)$ such that $A^*(\zeta) = v - G_h(v)$. Recall that $V := H^1_0(D)$ is equipped with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} = |v|_{H^1(D)}$. We infer that

$$\begin{aligned}\|v - G_h(v)\|^2_{L^2(D)} &= a(v - G_h(v), \zeta) = a(v - G_h(v), \zeta - \mathcal{I}^{\mathrm{g,av}}_{h0}(\zeta))\\ &\le \|a\|\,\|\nabla(v - G_h(v))\|_{\boldsymbol{L}^2(D)}\|\nabla(\zeta - \mathcal{I}^{\mathrm{g,av}}_{h0}(\zeta))\|_{\boldsymbol{L}^2(D)}\\ &\le c\,\|a\|h^r|v|_{H^{1+r}(D)}h^s\ell_D^{-1-s}\|\zeta\|_{H^{1+s}(D)},\end{aligned}$$

and the assertion follows since $\|\zeta\|_{H^{1+s}(D)} \le c\,\alpha^{-1}\ell_D^2\|v - G_h(v)\|_{L^2(D)}$.

Exercise 32.2 (H^{-1}-estimate). Let $z \in H^1(D)$. Let $\zeta \in H^1_0(D)$ be such that $A^*(\zeta) = z$. Recall that $V := H^1_0(D)$ is equipped with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} = |v|_{H^1(D)}$. Since exact adjoint consistency holds true, we infer that

$$\begin{aligned}(u - u_h, z)_{L^2(D)} &= a(u - u_h, \zeta) = a(u - u_h, \zeta - w_h)\\ &\le \|a\|\,\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)}\|\nabla(\zeta - w_h)\|_{\boldsymbol{L}^2(D)},\end{aligned}$$

for all $w_h \in V_h$. Since $k \ge 1 + s$, we infer that

$$\begin{aligned}\inf_{w_h\in V_h}\|\nabla(\zeta - w_h)\|_{\boldsymbol{L}^2(D)} &\le c\,h^{1+s}|\zeta|_{H^{2+s}(D)}\\ &\le c\,h^{1+s}\ell_D^{-2-s}\|\zeta\|_{H^{2+s}(D)}\\ &\le c\,h^{1+s}c_{\mathrm{smo}}\alpha^{-1}\ell_D^{-s}\|z\|_{H^1(D)}\\ &\le c'\,h^{1+s}c_{\mathrm{smo}}\alpha^{-1}\ell_D^{1-s}\|\nabla z\|_{\boldsymbol{L}^2(D)},\end{aligned}$$

where the last bound follows from the Poincaré–Steklov inequality. Combining the two bounds, dividing by $\|\nabla z\|_{\boldsymbol{L}^2(D)}$, and taking the supremum over $z \in H^1_0(D)$ leads to the expected estimate.

Exercise 32.3 (Compactness). (i) We have

$$\begin{aligned}\|G_h - I\|_{\mathcal{L}(V;L)} &= \sup_{v\in V} \frac{\|G_h(v)-v\|_L}{\|v\|_V} = \sup_{v\in V\setminus V_h} \frac{\|G_h(v)-v\|_L}{\|v\|_V} \\ &= \sup_{v\in V\setminus V_h} \frac{\|G_h(v)-v\|_L}{\|G_h(v)-v\|_V}\frac{\|G_h(v)-v\|_V}{\|v\|_V} \\ &\le \epsilon(h) \sup_{v\in V\setminus V_h} \frac{\|G_h(v)-v\|_V}{\|v\|_V}.\end{aligned}$$

Using the error estimate (26.18), we obtain

$$\|G_h - I\|_{\mathcal{L}(V;L)} \le \frac{\|a\|_{V\times V}}{\alpha_h}\epsilon(h) \sup_{u\in V\setminus V_h} \inf_{v_h\in V_h} \frac{\|u-v_h\|_V}{\|u\|_V} \le \frac{\|a\|}{\alpha}\epsilon(h).$$

(ii) Using Step (i) and $\lim_{h\to 0}\epsilon(h) = 0$, we infer that $\lim_{h\to 0}\|G_h - I\|_{\mathcal{L}(V;L)} = 0$. But G_h is compact since its rank is finite (recall that V_h is finite-dimensional). Hence, I is compact (see Theorem A.21).
(iii) Let us assume that I is compact. Let $y \in V$ and $f \in L$. By definition, $I(y) = y$ and

$$\begin{aligned}\langle A^*(R(f)), y\rangle_{V',V} &= \overline{\langle A(y), R(f)\rangle_{V',V}} = \overline{a(y,R(f))} = \overline{(y,f)_L} \\ &= \overline{(I(y),f)_L} = \langle I^*(f), y\rangle_{V',V},\end{aligned}$$

which proves that $A^*\circ R = I^*$. Since A is an isomorphism, so is A^*, whence we infer that $R = (A^*)^{-1}\circ I^*$. Schauder's theorem (Theorem C.48) implies that $I^* : V' \to L' \equiv L$ is compact, which, in turn, proves that $R = (A^*)^{-1}\circ I^*$ is compact.
(iv) Let $R_h : L \to V_h$ be the operator defined by $a(v_h, R_h(f)) := (v_h, f)_L$ for all $v_h \in V_h$ and all $f \in L$. Let $f \in L$. The error estimate (26.18) for the adjoint problem gives

$$\|R(f) - R_h(f)\|_V \le \frac{\|a\|}{\alpha}\inf_{w_h\in V_h}\|R(f)-w_h\|_V \le \frac{\|a\|}{\alpha}\|R(f) - P_h^V(R(f))\|_V,$$

where we used that the stability constant for the discrete adjoint problem is again α (see Remark 26.8) together with the property $\inf_{w_h\in V_h}\|R(f)-w_h\|_V = \|R(f)-P_h^V(R(f))\|_V$.
(v) Since we assume that I is compact, we know from Step (iii) that R is also compact. Let B_L be the unit ball in V and $Z := R(B_L)$. Since R is compact, for every $\epsilon > 0$ there is a finite set of points $\{x_i\}_{i\in I}$ in $Z \subset V$ such that for all $v \in Z$, there is $i \in I$ such that $\|v - x_i\|_V \le \epsilon$. Let $f \in B_L$. There is $i \in I$ s.t. $\|R(f) - x_i\|_V \le \epsilon$ and

$$\begin{aligned}\|R(f) - P_h^V(R(f))\|_V &\le \|R(f)-x_i\|_V + \|x_i - P_h^V(x_i)\|_V + \|P_h^V(x_i - R(f))\|_V \\ &\le 2\epsilon + \|x_i - P_h^V(x_i)\|_V.\end{aligned}$$

Hence, we have

$$\|R - P_h^V\circ R\|_{\mathcal{L}(L;V)} = \sup_{f\in B_L}\|R(f) - P_h^V(R(f))\|_V \le 2\epsilon + \max_{i\in I}\|x_i - P_h^V(x_i)\|_V.$$

Using that $\lim_{h\to 0}\|x_i - P_h^V(x_i)\|_V = 0$ for all $i \in I$ (which a consequence of the approximability assumption), and recalling that $\mathrm{card}(I)$ is finite, we infer that

$$\lim_{h\to 0}\max_{i\in I}\|x_i - P_h^V(x_i)\|_V = \max_{i\in I}\lim_{h\to 0}\|x_i - P_h^V(x_i)\|_V = 0.$$

As a result, we have $\lim_{h\to 0}\|R - P_h^V \circ R\|_{\mathcal{L}(L;V)} \le 2\epsilon$. Since ϵ is arbitrary, we conclude that

$$\lim_{h\to 0}\|R - P_h^V \circ R\|_{\mathcal{L}(L;V)} = 0.$$

Then Step (iv) implies that

$$\lim_{h\to 0}\|R - R_h\|_{\mathcal{L}(L;V)} = 0.$$

(vi) We now estimate $\epsilon(h)$. Let $v \in V\backslash V_h$, i.e., $v - G_h(v) \neq 0$. We observe that

$$\begin{aligned}
\|v - G_h(v)\|_L &= \sup_{f\in B_L} |(v - G_h(v), f)_L| = \sup_{f\in B_L} |a(v - G_h(v), R(f))_L| \\
&= \sup_{f\in B_L} |a(v - G_h(v), R(f) - R_h(f))_L| \\
&\le \|a\|\,\|v - G_h(v)\|_V \sup_{f\in B_L} \|R(f) - R_h(f)\|_V \\
&= \|a\|\,\|v - G_h(v)\|_V \|R - R_h\|_{\mathcal{L}(L;V)}.
\end{aligned}$$

We infer that $\epsilon(h) \le \|a\|\|R - R_h\|_{\mathcal{L}(L;V)}$, and the conclusion follows from Step (v).

Exercise 32.4 (Source approximation). (i) Either we directly invoke Strang's first lemma or we redo the argument from the proof of Lemma 27.5. We follows here the second option. For all $v_h \in V_h$, we have

$$\begin{aligned}
\|u - u_h\|_V &\le \|u - v_h\|_V + \|v_h - u_h\|_V \\
&\le \|u - v_h\|_V + \frac{1}{\alpha}\sup_{w_h\in V_h} \frac{a(v_h - u_h, w_h)}{\|w_h\|_V} \\
&\le \|u - v_h\|_V + \frac{1}{\alpha}\sup_{w_h\in V_h} \frac{a(v_h, w_h) - \ell_h(w_h)}{\|w_h\|_V} \\
&\le \|u - v_h\|_V + \frac{1}{\alpha}\sup_{w_h\in V_h} \frac{a(v_h, w_h) - \ell(w_h) + \ell(w_h) - \ell_h(w_h)}{\|w_h\|_V} \\
&\le \|u - v_h\|_V + \frac{1}{\alpha}\sup_{w_h\in V_h} \frac{a(v_h - u, w_h) + \ell(w_h) - \ell_h(w_h)}{\|w_h\|_V} \\
&\le \|u - v_h\|_V + \frac{\|a\|}{\alpha}\|u - v_h\|_V + \delta_h,
\end{aligned}$$

with $\delta_h := \frac{1}{\alpha}\sup_{w_h\in V_h} \frac{\int_D (f - \mathcal{I}_h^{\rm b}(f))w_h\,{\rm d}x}{\|w_h\|_V}$. Recalling that $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} = |v|_{H^1(D)}$, δ_h is bounded as

$$\begin{aligned}
\delta_h &= \frac{1}{\alpha}\sup_{w_h\in V_h}\inf_{v_h\in P_{k'}^{\rm b}(\mathcal{T}_h)} \frac{\int_D (f - \mathcal{I}_h^{\rm b}(f))(w_h - v_h)\,{\rm d}x}{\|w_h\|_V} \\
&\le \frac{1}{\alpha}\|f - \mathcal{I}_h^{\rm b}(f)\|_{L^2(D)} \sup_{w_h\in V_h} \frac{ch\|\nabla w_h\|_{\boldsymbol{L}^2(D)}}{\|w_h\|_V} \\
&\le c\,\alpha^{-1}h\|f - \mathcal{I}_h^{\rm b}(f)\|_{L^2(D)}.
\end{aligned}$$

This means that

$$\|u - u_h\|_V \le \left(1 + \frac{\|a\|}{\alpha}\right)\inf_{v\in V_h}\|u - v_h\|_V + c\,\alpha^{-1}h\|f - \mathcal{I}_h^{\rm b}(f)\|_{L^2(D)}.$$

If $u \in H^{k+1}(D)$ and $f \in H^{k'}(D)$, then

$$\|u - u_h\|_V \le c\,(h^k |u|_{H^{k+1}(D)} + \alpha^{-1} h^{1+k'} |f|_{H^{k'}(D)}).$$

So it suffices that $k' := k-1$ to obtain optimality in the H^1-norm.
(ii) We now reformulate (32.19). Let $\zeta_{u-u_h} \in V$ solve $a(v, \zeta_{u-u_h}) = (v, u-u_h)_{L^2(D)}$ for all $v \in V$. Elliptic regularity implies that $\|\zeta_{u-u_h}\|_{H^{1+s}(D)} \le c_{\rm smo}\alpha^{-1}\ell_D^2 \|u-u_h\|_{L^2(D)}$. For all $v_h \in V_h$, we have

$$\begin{aligned}
\|u-u_h\|_{L^2(D)}^2 &= a(u-u_h, \zeta_{u-u_h}) \\
&= a(u-u_h, \zeta_{u-u_h} - v_h) + a(u, v_h) - a(u_h, v_h) \\
&= a(u-u_h, \zeta_{u-u_h} - v_h) + \int_D (f - \mathcal{I}_h^{\rm b}(f)) v_h \,{\rm d}x \\
&= a(u-u_h, \zeta_{u-u_h} - v_h) + \inf_{w_h \in P_{k'}^{\rm b}(\mathcal{T}_h)} \int_D (f - \mathcal{I}_h^{\rm b}(f))(v_h - w_h)\,{\rm d}x.
\end{aligned}$$

Let us take v_h to be the best approximation of ζ_{u-u_h} in V_h in the V-norm. Since

$$\inf_{w_h \in P_{k'}^{\rm b}(\mathcal{T}_h)} \|v_h - w_h\|_{L^2(D)} \le c' h \|\nabla v_h\|_{\boldsymbol{L}^2(D)} \le c'' h \|\nabla \zeta_{u-u_h}\|_{\boldsymbol{L}^2(D)},$$

we infer that

$$\begin{aligned}
\|u-u_h\|_{L^2(D)}^2 &\le \|a\| \|u-u_h\|_V c h^s |\zeta_{u-u_h}|_{H^{1+s}(D)} + \|f - \mathcal{I}_h^{\rm b}(f)\|_{L^2(D)} c'' h \|\nabla \zeta_{u-u_h}\|_{\boldsymbol{L}^2(D)} \\
&\le c\left(h^s \ell_D^{1-s} \|u-u_h\|_V + \alpha^{-1} h \ell_D \|f - \mathcal{I}_h^{\rm b}(f)\|_{L^2(D)}\right) \|u-u_h\|_{L^2(D)}.
\end{aligned}$$

(Note that we have hidden the nondimensional factor $\frac{\|a\|}{\alpha}$ in the generic constant c.) If $u \in H^{k+1}(D)$ and $f \in H^{k'}(D)$, and assuming $s=1$, we obtain

$$\|u-u_h\|_{L^2(D)} \le c\left(h^{k+1} |u|_{H^{k+1}(D)} + \alpha^{-1} h^{k'+1} \ell_D |f|_{H^{k'}(D)}\right),$$

where we used the bound on $\|u-u_h\|_V$ from the previous step and $h \le \ell_D$. We now obtain optimality in the L^2-norm if $k' := k$.

Exercise 32.5 (Advection-diffusion, 1D). (i) The stiffness matrix is given by

$$\mathcal{A} = \frac{\nu}{h} \mathrm{tridiag}\Big(-1 - \frac{\gamma}{2}, 2, -1 + \frac{\gamma}{2}\Big).$$

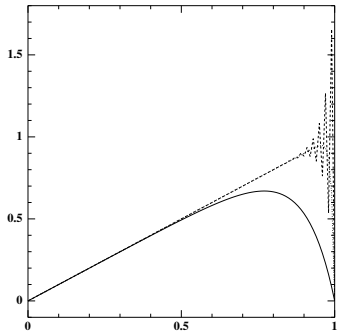
(ii) Assuming that $f = 1$, the linear system to be solved is $\mathcal{A}\mathsf{U} = h(1, \dots, 1)^\mathsf{T}$. Since $\mathcal{A}\mathsf{U}^0 = (h, \dots, h, h + \gamma^{-1}(1 - \frac{\gamma}{2}))^\mathsf{T}$ (observe that $h(I+1) = 1$), we infer that $\mathcal{A}\tilde{\mathsf{U}} = (0, \dots, 0, \gamma^{-1}(\frac{\gamma}{2} - 1))^\mathsf{T}$. If $\gamma = 2$, then $\tilde{\mathsf{U}} = 0$. Let us now assume that $\gamma \ne 2$. Using $\tilde{\mathsf{U}}_i = \varrho + \theta \delta^i$, we infer from the rows $\{2{:}I{-}1\}$ of the linear system that

$$\left(-1 - \tfrac{\gamma}{2}\right) + 2\delta + \left(-1 + \tfrac{\gamma}{2}\right)\delta^2 = 0,$$

so that $\delta = 1$ or $\delta = \frac{2+\gamma}{2-\gamma}$. The first row of the system yields $\theta = -\varrho$. From the last row of the system, we finally infer that $\frac{\nu}{h}(1 - \frac{\gamma}{2})\varrho(1 - \delta^{I+1}) = \gamma^{-1}(\frac{\gamma}{2} - 1)$, i.e., $b\varrho(1 - \delta^{I+1}) = -1$. Notice that $\delta \ne 1$ because we assumed that $\gamma = \frac{bh}{\nu} \ne 0$. Hence, $-\theta = \varrho = -b^{-1}(1 - \delta^{I+1})^{-1}$, that is,

$$\tilde{\mathsf{U}}_i = -b^{-1} \frac{\delta^i - 1}{\delta^{I+1} - 1}, \qquad \delta = \frac{2+\gamma}{2-\gamma}.$$

When $\gamma > 2$, the components of the vector $\tilde{\mathsf{U}}$ oscillate between positive and negative values. The approximate solutions for $\gamma \in \{0.1, 1, 10\}$ obtained with $h := 10^{-2}$ are plotted on the figure shown here. We observe that for $\gamma = 10$ the approximate solution exhibits spurious oscillations close to the boundary layer. Instead, the approximate solutions for $\gamma := 1$ and $\gamma := 0.1$ match well the exact solution.



(iii) Setting $V := \{v \in H^1(D) \mid v(0) = 0\}$, the weak formulation now consists of seeking $u \in V$ such that $a(u, w) = \ell(w)$ for all $w \in V$. Since $\int_0^1 bv'v \, \mathrm{d}x = \frac{1}{2}bv(1)^2 \geq 0$, the bilinear form a is still coercive on V. The stiffness matrix is of order $(I + 1)$ and has the following tridiagonal structure:

$$\mathcal{A} = \frac{\nu}{h} \begin{pmatrix} c_0 & c_+ & 0 & \dots & 0 \\ c_- & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & c_0 & c_+ \\ 0 & \dots & 0 & c_- & c_0' \end{pmatrix},$$

with $c_0 := 2$, $c_0' := 1 + \frac{\gamma}{2}$, $c_+ := -1 + \frac{\gamma}{2}$, and $c_- := -1 - \frac{\gamma}{2}$. We infer that $(\nu + \frac{bh}{2})(\mathsf{U}_{I+1} - \mathsf{U}_I) = \int_{x_N}^{x_{I+1}} f\varphi_{I+1} \, \mathrm{d}x$, so that

$$\frac{\mathsf{U}_{I+1} - \mathsf{U}_I}{h} = \frac{2\int_{x_I}^{x_{I+1}} f\varphi_{I+1} \, \mathrm{d}x}{2\nu + bh}.$$

Hence, $\frac{\mathsf{U}_{I+1} - \mathsf{U}_I}{h} \to 0$ as $h \to 0$ with fixed $\nu > 0$, whereas $\frac{\mathsf{U}_{I+1} - \mathsf{U}_I}{h} \to \frac{f(1)}{b}$ as $\nu \to 0$ with fixed $h \in \mathcal{H}$.

Chapter 33

H^1-conforming approximation (II)

Exercises

Exercise 33.1 (Regularity assumption). Let u_h solve (33.5). Assume that $u \in H^{1+r}(D)$ with $r \in (0,k]$. Prove that $\|u-u_h\|_{H^1(D)} \le c(h^r|u|_{H^{1+r}(D)} + (\sum_{F\in\mathcal{F}_h^\partial} h_F^{-1}\|g-g_h\|_{L^2(F)}^2)^{\frac12})$. (*Hint*: consider $v_h := \mathcal{I}_{h0}^{\mathrm{g,av}}(u) + \sum_{a\in\mathcal{A}_h^\partial}\sigma_a^\partial(g)\varphi_a$, and follow the proof of Theorem 22.14 to bound $\|u-v_h\|_{H^1(D)}$.)

Exercise 33.2 (Non-homogeneous Dirichlet). Let $\mathcal{A}$ denote the system matrix in (33.10). Let $\mathsf{R}\in\mathbb{R}^I$ and let $k\ge 1$. Consider the Krylov space $S_k := \mathrm{span}\{\mathsf{R},\mathcal{A}\mathsf{R},\dots,\mathcal{A}^{k-1}\mathsf{R}\}$. For all $\mathsf{V}\in\mathbb{R}^I$, write $\mathsf{V} := (\mathsf{V}^\circ,\mathsf{V}^\partial)^\intercal$. Assume that $\mathsf{R}^\partial = 0$. (i) Prove that $\mathsf{Y}^\partial = 0$ for all $\mathsf{Y}\in S_k$. (ii) Prove that if $\mathcal{A}^{\circ\circ}$ is symmetric, the restriction of $\mathcal{A}$ to S_k is symmetric.

Exercise 33.3 (DMP). Assume that the stiffness matrix is a Z-matrix. Assume the following: (i) $\mathcal{A}_{ii} \ge -\sum_{j\neq i}\mathcal{A}_{ij}$ for all $i\in\{1{:}I\}$; (ii) $\exists i_*\in\{1{:}I\}$ such that $\mathcal{A}_{i_*i_*} > -\sum_{j\neq i_*}\mathcal{A}_{i_*j}$; (iii) For all $i\in\{1{:}I\}$, $i\neq i_*$, there exists a path $[i =: i_1,\dots,i_J := i_*]$ such that $\mathcal{A}_{i_ji_{j+1}} < 0$ for all $j\in\{1{:}J{-}1\}$. Prove that $\mathcal{A}$ is a nonsingular M-matrix. (*Hint*: let $\mathsf{B}\le 0$, let $\mathsf{U} := \mathcal{A}^{-1}\mathsf{B}$, and proceeding by contradiction, assume that there is $i\in\{1{:}I\}$ s.t. $U_i = \max_{j\in\{1:I\}}U_j > 0$.)

Exercise 33.4 (Obtuse mesh). The mesh shown in Figure 33.1 contains three interior nodes with coordinates $\boldsymbol{z}_1 := (1,1)$, $\boldsymbol{z}_2 := (3,1)$, and $\boldsymbol{z}_3 := (2,\frac32)$. The sum of the two angles opposite the edge linking $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ is larger than π. (i) Assemble the 3×3 stiffness matrix $\mathcal{A}$ generated by the three shape functions associated with the three interior nodes $\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_3$. Is $\mathcal{A}$ a Z-matrix? (*Hint*: the local stiffness matrix is translation- and scale-invariant, there are four shapes of triangles in the mesh, and one can work on triangles with vertices $((0,0),(1,0),(0,1))$, $((0,0),(1,0),(0,\frac12))$, $((-1,0),(1,0),(0,\frac12))$, and $((-1,1),(1,1),(0,1))$.) (ii) Compute $\mathcal{A}^{-1}$. Is $\mathcal{A}$ an M-matrix?

Exercise 33.5 (1D DMP). Consider the equation $\mu u+\beta u'-\nu u'' = f$ in $D := (0,1)$. Let $\mathcal{T}_h$ be the uniform mesh composed of the cells $[ih,(i{+}1)h]$, $\forall i\in\{0{:}I\}$, with uniform meshsize $h := \frac{1}{I+1}$. Assume $\mu\in\mathbb{R}_+$, $\beta\in\mathbb{R}$, $\nu\in\mathbb{R}_+$ and $f\in L^1(D)$. Let $u_h := \sum_{i\in\{0:I+1\}}U_i\varphi_i \in P_1^{\mathrm{g}}(\mathcal{T}_h)$ be such that $\int_D((\mu u_h+\beta u_h')\varphi_i+\nu u_h'\varphi_i')\,\mathrm{d}x = \int_D f\varphi_i\,\mathrm{d}x$ for all $i\in\{1{:}I\}$. Let $F_i := \int_D f\varphi_i\,\mathrm{d}x/\int_D\varphi_i\,\mathrm{d}x$. Assume that $\frac{\nu}{h}\ge\frac{|\beta|}{2}+\frac{\mu h}{6}$. (i) Show that $\min(U_{i-1},U_{i+1},\frac{F_i}{\mu}) \le U_i \le \max(U_{i-1},U_{i+1},\frac{F_i}{\mu})$ for all $i\in\{1{:}I\}$. (*Hint*: write the linear system as $\mu U_i+\alpha_{i-1}(\mu,\beta,\nu)(U_i-U_{i-1})+\alpha_{i+1}(\mu,\beta,\nu)(U_i-U_{i+1}) = F_i$.) (ii) Show that $\min(U_0,U_{I+1},\frac{\min_{j\in\{1:I\}}F_j}{\mu}) \le U_i \le \max(U_0,U_{I+1},\frac{\max_{j\in\{1:I\}}F_j}{\mu})$ for all $i\in\{1{:}I\}$.

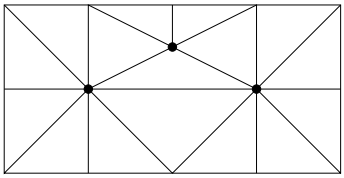


Figure 33.1: Illustration for Exercise 33.4.

Exercise 33.6 (1D DMP, pure diffusion). Let $D := (0,1)$, $f \in L^\infty(D)$, and a nonuniform mesh $\mathcal{T}_h$ of D with nodes $\{x_i\}_{i\in\{0:I+1\}}$. Let $u_h \in P_1^{\rm g}(\mathcal{T}_h)$ be s.t. $u_h(0) = a$, $u_h(1) = b$, and $\int_D u_h' v_h' \,\mathrm{d}x = \int_D f v_h \,\mathrm{d}x$ for all $v_h \in P_{1,0}^{\rm g}(\mathcal{T}_h)$. (i) Show that $\max_{x\in D} u_h(x) \le \max(a,b) + \frac{1}{4}\operatorname{ess\,sup}_{x\in D} f(x)$. (*Hint*: test with $\phi_h \in P_{1,0}^{\rm g}(\mathcal{T}_h)$ s.t. $\phi_{h|[0,x_i]} := \frac{x}{x_i}$ and $\phi_{h|[x_i,1]} := \frac{1-x}{1-x_i}$ for all $i \in \{1{:}I\}$.) (ii) Let ϕ_h be the function defined in the hint. Compute $-\partial_{xx}\phi_h$. Comment on the result.

Exercise 33.7 (Maximum principle). Let D be a bounded Lipschitz domain in $\mathbb{R}^d$. Let $\boldsymbol{x}_0 \in D$ and $R \in \mathbb{R}$ be s.t. $\max_{\boldsymbol{x}\in D} \|\boldsymbol{x} - \boldsymbol{x}_0\|_{\ell^2} \le R$. (i) Let $\phi(\boldsymbol{x}) := -\frac{1}{2d}\|\boldsymbol{x} - \boldsymbol{x}_0\|_{\ell^2}^2$. Compute $-\Delta\phi$. Give an upper bound on $\max_{\boldsymbol{x}\in D}\phi(\boldsymbol{x})$ and a lower bound on $\min_{\boldsymbol{x}\in\partial D}\phi(\boldsymbol{x})$. (ii) Let $f \in L^\infty(D)$ and let $u \in H^1(D)$ solve $-\Delta u = f$. Let $M := \operatorname{ess\,sup}_{\boldsymbol{x}\in D} f(\boldsymbol{x})$. Give an upper bound on $-\Delta(u - M\phi)$. (iii) Prove that $\max_{\boldsymbol{x}\in D} u(\boldsymbol{x}) \le \max_{\boldsymbol{x}\in\partial D} u(\boldsymbol{x}) + M_+\frac{R^2}{2d}$ with $M_+ := \max(M,0)$. (*Hint*: use (i) from Theorem 33.6.)

Solution to exercises

Exercise 33.1 (Regularity assumption). Let us set

$$v_h := \mathcal{I}_{h0}^{\rm g,av}(u) + u_{gh}, \qquad u_{gh} := \sum_{a\in\mathcal{A}_h^\partial} \sigma_a^\partial(g)\varphi_a.$$

Observing that $v_{h|\partial D} = \sum_{a\in\mathcal{A}_h^\partial}\sigma_a^\partial(g)\varphi_{a|\partial D} = g_h$, we infer that $u_h - v_h \in V_h$. Proceeding as in the proof of Theorem 33.2, we infer that $\|u - u_h\|_{H^1(D)} \le c\|u - v_h\|_{H^1(D)}$. Since $\mathcal{J}_{h0}^{\rm g,av}(u_{gh}) = 0$ and recalling that $\mathcal{I}_{h0}^{\rm g,av}(u) = \mathcal{J}_{h0}^{\rm g,av}\mathcal{I}_h^{\rm g,\sharp}(u)$, we infer that

$$u - v_h = (u - \mathcal{I}_h^{\rm g,\sharp}(u)) + (w_h - \mathcal{J}_{h0}^{\rm g,av}(w_h)) =: \mathfrak{T}_1 + \mathfrak{T}_2,$$

with $w_h := \mathcal{I}_h^{\rm g,\sharp}(u) - u_{gh}$. Owing to Theorem 18.14, we infer that $\|\mathfrak{T}_1\|_{H^1(D)} \le ch^r|u|_{H^{1+r}(D)}$. Concerning $\mathfrak{T}_2$, we infer from the proof of Theorem 22.14 (with $m := 1$ and $p := 2$) that

$$\|\mathfrak{T}_2\|_{H^1(K)} \le c\sum_{F\in\mathcal{F}_K^\circ} h_K^{-\frac12}\|[\![w_h]\!]_F\|_{L^2(F)} + c'\sum_{F\in\mathcal{F}_K^\partial} h_K^{-\frac12}\|w_h\|_{L^2(F)}.$$

Since $[\![u_{gh}]\!]_F = 0$, the first sum on the right-hand side is bounded as before. For the second one, we write $\|w_h\|_{L^2(F)} \le \|u - \mathcal{I}_h^{\rm g,\sharp}(u)\|_{L^2(F)} + \|u - u_{gh}\|_{L^2(F)}$ by the triangle inequality, and $\|u - \mathcal{I}_h^{\rm g,\sharp}(u)\|_{L^2(F)}$ is bounded as before using a multiplicative trace inequality. Finally, we observe that $(u - u_{gh})_{|F} = g - g_h$ for all $F \in \mathcal{F}_h^\partial$, and we invoke the regularity of the mesh sequence to replace h_K by h_F.

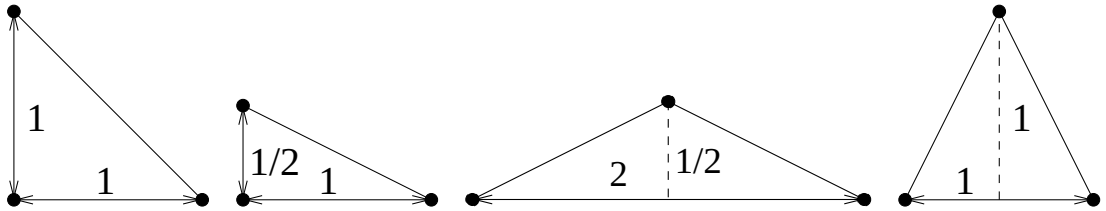
Exercise 33.2 (Non-homogeneous Dirichlet). (i) A direct computation shows that $\mathcal{A}\mathsf{R} = (\mathcal{A}^{\circ\circ}\mathsf{R}^\circ + \mathcal{A}^{\circ\partial}\mathsf{R}^\partial, \mathsf{R}^\partial)^\mathsf{T} = (\mathcal{A}^{\circ\circ}\mathsf{R}^\circ, 0)^\mathsf{T}$ since $\mathsf{R}^\partial = 0$. By induction, we infer that $(\mathcal{A}^l\mathsf{R})^\partial = 0$ for all $l \geq 0$.
(ii) Let X and Y be two vectors in the Krylov subspace S_k. Owing to Step (i), we infer that $\mathsf{X}^\partial = \mathsf{Y}^\partial = 0$. As a result, we have $(\mathcal{A}\mathsf{X}, \mathsf{Y})_{\ell^2(\mathbb{R}^I)} = (\mathcal{A}^{\circ\circ}\mathsf{X}^\circ, \mathsf{Y}^\circ)_{\ell^2(\mathbb{R}^{I^\circ})}$ with $I = \mathrm{card}(\mathcal{A}_h)$ and $I^\circ = \mathrm{card}(\mathcal{A}_h^\circ)$. Hence, the restriction of $\mathcal{A}$ to S_k has the same symmetry properties as $\mathcal{A}^{\circ\circ}$.

Exercise 33.3 (DMP). Let $\mathsf{B} \leq 0$ and let $\mathsf{U} := \mathcal{A}^{-1}\mathsf{B}$. Proceeding by contradiction, assume that there is $i \in \{1{:}I\}$ such that $U_i = \max_{j\in\{1:I\}} U_j > 0$. Since $\mathcal{A}\mathsf{U} = \mathsf{B}$, we infer that

$$0 \geq B_i = \mathcal{A}_{ii}U_i + \sum_{j\neq i} \mathcal{A}_{ij}U_j \geq \Delta_i U_i + \sum_{j\neq i} \mathcal{A}_{ij}(U_j - U_i),$$

where $\Delta_i := \mathcal{A}_{ii} - \sum_{j\neq i}\mathcal{A}_{ij} \geq 0$ owing to Assumption (i), whereas the second term on the right-hand side is nonnegative since $\mathcal{A}$ is a Z-matrix and $U_i = \max_{j\in\{1:I\}} U_j$. Hence, both addends vanish. As a result, $i \neq i_*$ owing to Assumption (ii). Exploiting Assumption (iii), we consider the path $[i =: i_1, \dots, i_J := i_*]$ such that $\mathcal{A}_{i_j i_{j+1}} < 0$ for all $j \in \{1{:}J{-}1\}$. Since we already know that $\mathcal{A}_{ij}(U_j - U_i) = 0$ for all $j \neq i$, we infer that $U_{i_2} = U_i$. Reasoning similarly, we infer that $U_{i_j} = U_i$ for all $j \in \{1{:}J\}$, which provides the expected contradiction once we reach $i_J = i_*$.

Exercise 33.4 (Obtuse mesh). (i) Let us work on the following triangles:



In all the cases, the vertices are numbered anticlockwise starting from the lower left vertex. The local stiffness matrices are, respectively,

$$\frac{1}{2}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \frac{1}{4}\begin{pmatrix} 5 & -1 & -4 \\ -1 & 1 & 0 \\ -4 & 0 & 4 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} \frac{5}{4} & \frac{3}{4} & -2 \\ \frac{3}{4} & \frac{5}{4} & -2 \\ -2 & -2 & 4 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} \frac{5}{4} & -\frac{3}{4} & -\frac{1}{2} \\ -\frac{3}{4} & \frac{5}{4} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}.$$

(Observe that the row- and columnwise sums of the above matrices vanish.) The entries of the stiffness matrix are such that $\mathcal{A}_{11} = 6\frac{1}{2} + \frac{5}{8} + \frac{5}{4} = \frac{17}{4}$, $\mathcal{A}_{12} = 0 + \frac{3}{8} = \frac{3}{8}$, $\mathcal{A}_{13} = -1 - \frac{1}{4} = -\frac{5}{4}$, $\mathcal{A}_{33} = 2 + 2\frac{1}{2} + 2 = 5$, and the other entries are evaluated using symmetries so that

$$\mathcal{A} = \begin{pmatrix} \frac{17}{4} & \frac{3}{8} & -\frac{5}{4} \\ \frac{3}{8} & \frac{17}{4} & -\frac{5}{4} \\ -\frac{5}{4} & -\frac{5}{4} & 5 \end{pmatrix}.$$

Hence, $\mathcal{A}$ is not a Z-matrix.
(ii) Computing the inverse of $\mathcal{A}$, we obtain

$$\mathcal{A}^{-1} = \begin{pmatrix} \frac{63}{248} & -\frac{1}{248} & \frac{1}{16} \\ -\frac{1}{248} & \frac{63}{248} & \frac{1}{16} \\ \frac{1}{16} & \frac{1}{16} & \frac{37}{160} \end{pmatrix}.$$

Hence, $\mathcal{A}$ is not an M-matrix.

Exercise 33.5 (1D DMP). (i) The discrete system is written

$$\mu\frac{h}{6}(U_{i-1}+4U_i+U_{i+1})+\frac{\beta}{2}(U_i-U_{i-1})+\frac{\beta}{2}(U_{i+1}-U_i) \\ +\frac{\nu}{h}(U_i-U_{i-1})+\frac{\nu}{h}(U_i-U_{i+1})=hF_i.$$

The contribution from the mass matrix can be rewritten

$$\mu\frac{h}{6}(U_{i-1}+4U_i+U_{i+1})=\mu hU_i-\mu\frac{h}{6}(U_i-U_{i-1})-\mu\frac{h}{6}(U_i-U_{i+1}).$$

In conclusion, we have

$$\mu hU_i+(U_i-U_{i-1})\Big(-\frac{\mu h}{6}+\frac{\beta}{2}+\frac{\nu}{h}\Big)+(U_i-U_{i+1})\Big(-\frac{\mu h}{6}-\frac{\beta}{2}+\frac{\nu}{h}\Big)=hF_i.$$

Assume first that $U_i\le\max(U_{i-1},U_{i+1})$, so that $U_i\le\max(U_{i-1},U_{i+1},\frac{F_i}{\mu})$. Assume now that $U_i>\max(U_{i-1},U_{i+1})$, so that the above identity and the assumption $\frac{\nu}{h}\ge\frac{|\beta|}{2}+\frac{\mu h}{6}$ imply that

$$\mu hU_i\le hF_i,$$

which means that $U_i\le\frac{F_i}{\mu}$. Thus, we infer that $U_i\le\max(U_{i-1},U_{i+1},\frac{F_i}{\mu})$. The other inequality is proved similarly.
(ii) By induction, we have $U_i\le\max(U_{i-1},U_{i+l},\max_{j\in\{i:i+l-1\}}\frac{F_j}{\mu})$ for all $l\in\{1{:}I+1-i\}$. Hence, $U_i\le\max(U_{i-1},U_{I+1},\max_{j\in\{i:I\}}\frac{F_j}{\mu})$. Similarly, we have $U_i\le\max(U_{i-l},U_{I+1},\max_{j\in\{i-l+1:I\}}\frac{F_j}{\mu})$ for all $l\in\{1{:}i\}$. Hence,

$$U_i\le\max(U_0,U_{I+1},\max_{j\in\{1:I\}}\frac{F_j}{\mu}).$$

The other inequality is proved similarly.

Exercise 33.6 (1D DMP, pure diffusion). (i) If u_h is maximum at $x_0=0$ or at $x_{I+1}=1$, there is nothing to prove since $\max_{x\in D}u_h(x)=\max(a,b)\le\max(a,b)+\frac14\operatorname{ess\,sup}_{x\in D}f_+(x)$. Assume that u_h is maximum inside D. Since u_h is piecewise linear, the maximum must occur at a node x_i with $i\in\{1{:}I\}$. Let $\phi_h\in P^g_{1,0}(\mathcal{T}_h)$ be such that $\phi_{h|[0,x_i]}(x)=\frac{x}{x_i}$ and $\phi_{h|[x_i,1]}(x)=\frac{1-x}{1-x_i}$. Since $u_h(0)=a$, $u_h(1)=b$, and letting $\mathsf{U}_i:=u_h(x_i)$, we have

$$\begin{aligned}\int_D f\phi_h\,\mathrm{d}x=\int_D u_h'\phi_h'\,\mathrm{d}x&=\frac{1}{x_i}\int_0^{x_i}u_h'\,\mathrm{d}x-\frac{1}{1-x_i}\int_{x_i}^1u_h'\,\mathrm{d}x\\&=\frac{1}{x_i}(\mathsf{U}_i-a)-\frac{1}{1-x_i}(b-\mathsf{U}_i)\\&=\mathsf{U}_i(\frac{1}{x_i}+\frac{1}{1-x_i})-\frac{a}{x_i}-\frac{b}{1-x_i}\\&=\mathsf{U}_i\frac{1}{x_i(1-x_i)}-\frac{a(1-x_i)+bx_i}{x_i(1-x_i)}.\end{aligned}$$

We infer that

$$\begin{aligned}\mathsf{U}_i&=a(1-x_i)+bx_i+x_i(1-x_i)\int_D f\phi_h\,\mathrm{d}x\\&\le\max(a,b)+\Big(\operatorname*{ess\,sup}_{x\in D}f(x)\Big)\frac12\int_D\phi_h\,\mathrm{d}x\\&\le\max(a,b)+\frac14\operatorname*{ess\,sup}_{x\in D}f(x)\le\max(a,b)+\frac14\operatorname*{ess\,sup}_{x\in D}f_+(x).\end{aligned}$$

In conclusion, $\max_{x\in D} u_h(x) \le \max(a,b) + \frac{1}{4}\operatorname{ess\,sup}_{x\in D} f_+(x)$.
(ii) Let $\varphi \in C_0^\infty(D)$. We have

$$\begin{aligned}\int_D \phi_h\varphi''\,\mathrm{d}x = -\int_D \phi_h'\varphi'\,\mathrm{d}x &= -\frac{1}{x_i}\int_0^{x_i}\varphi'\,\mathrm{d}x + \frac{1}{1-x_i}\int_{x_i}^1 \varphi'\,\mathrm{d}x\\ &= -\varphi(x_i)(\frac{1}{x_i}+\frac{1}{1-x_i}) = -\frac{1}{x_i(1-x_i)}\langle\delta_{x_i},\varphi\rangle,\end{aligned}$$

where δ_{x_i} is the Dirac measure at x_i. Hence, $-\phi_h'' = \frac{1}{x_i(1-x_i)}\delta_{x_i}$. This means that $x_i(1-x_i)\phi_h$ is the Green function of the Laplace operator over $D := (0,1)$ with Dirichlet boundary conditions.

Exercise 33.7 (Maximum principle). (i) We have $-\Delta\phi = 1$ in D, $\max_{\boldsymbol{x}\in D}\phi(\boldsymbol{x}) = 0$, and $\min_{\boldsymbol{x}\in\partial D}\phi(\boldsymbol{x}) \ge -\frac{R^2}{2d}$.
(ii) The definitions give $-\Delta(u - M\phi) = f + M\Delta\phi = f - M \le 0$.
(iii) If $M \le 0$, then $f \le 0$, and using the hint, we infer that

$$\max_{\boldsymbol{x}\in D} u(\boldsymbol{x}) \le \max_{\boldsymbol{x}\in\partial D} u(\boldsymbol{x}) = \max_{\boldsymbol{x}\in\partial D} u(\boldsymbol{x}) + M_+\frac{R^2}{2d}.$$

Let us assume now that $M > 0$. Using the hint together with $-\Delta(u - M\phi) \le 0$, we infer that

$$\begin{aligned}\max_{\boldsymbol{x}\in D}(u - M\phi(\boldsymbol{x})) \le \max_{\boldsymbol{x}\in\partial D}(u - M\phi(\boldsymbol{x})) &\le \max_{\boldsymbol{x}\in\partial D} u(\boldsymbol{x}) + M\max_{\boldsymbol{x}\in\partial D} -\phi(\boldsymbol{x})\\ &\le \max_{\boldsymbol{x}\in\partial D} u(\boldsymbol{x}) - M\min_{\boldsymbol{x}\in\partial D}\phi(\boldsymbol{x}) \le \max_{\boldsymbol{x}\in\partial D} u(\boldsymbol{x}) + M\frac{R^2}{2d}.\end{aligned}$$

Using that $M > 0$ and $\phi \le 0$ gives

$$\max_{\boldsymbol{x}\in D} u(\boldsymbol{x}) \le \max_{\boldsymbol{x}\in D}(u(\boldsymbol{x}) - M\phi(\boldsymbol{x})).$$

Putting everything together, we conclude that

$$\max_{\boldsymbol{x}\in D} u(\boldsymbol{x}) \le \max_{\boldsymbol{x}\in\partial D} u(\boldsymbol{x}) + M_+\frac{R^2}{2d}.$$

Chapter 34

A posteriori error analysis

Exercises

Exercise 34.1 (Residual). Prove (34.10). (*Hint*: integrate by parts.)

Exercise 34.2 (Trace inequality in stars). Let $C_{\mathrm{tr},\boldsymbol{z}}$ be defined in (34.12). Prove that $C_{\mathrm{tr},\boldsymbol{z}} \le \varpi_{\boldsymbol{z}}^{\frac12}(dC_{\mathrm{PS},\boldsymbol{z}}^2 + 2C_{\mathrm{PS},\boldsymbol{z}})^{\frac12}$ with $\varpi_{\boldsymbol{z}} := h_{D_{\boldsymbol{z}}} \max_{F\in\mathcal{F}_{\boldsymbol{z}}^\circ} \frac{|F|}{|D_F|}$ and $D_F := \mathrm{int}(K_l \cup K_r)$ with $F := \partial K_l \cap \partial K_r$. (*Hint*: see the proof of Lemma 12.15.)

Exercise 34.3 (Bound on dual norm). (i) Prove that $\|T_K^{\mathrm{v}}(f)\|_{H^{-1}(K)} \le ch_K\|f\|_{L^2(K)}$ for all $f \in L^2(K)$. (*Hint*: use a scaled Poincaré–Steklov inequality for functions $\varphi \in H_0^1(K)$.) (ii) Prove that $\|T_F^{\mathrm{s}}(g)\|_{H^{-1}(D_F)} \le ch_F^{\frac12}\|g\|_{L^2(F)}$ for all $g \in L^2(F)$. (*Hint*: use the multiplicative trace inequality from Lemma 12.15.)

Exercise 34.4 (Oscillation). (i) Let $P_m^{(p)} : L^p(K) \to \mathbb{P}_m$ be the best-approximation operator in $L^p(K)$ for $p \in [1,\infty]$ and $m \in \mathbb{N}$. Prove that

$$\|(I - P_m^{(2)})(\theta v_h)\|_{L^2(K)} \le \|(I - P_{m-n}^{(\infty)})(\theta)\|_{L^\infty(K)}\|v_h\|_{L^2(K)},$$

for all $\theta \in L^\infty(K)$ and all $v_h \in \mathbb{P}_n$ with $n \le m$. (ii) Consider the oscillation indicators defined in (34.19) with $l^{\mathrm{v}} := 2k-2$ and $l^{\mathrm{s}} := 2k-1$. Prove that $\phi_K^{\mathrm{v}}(u_h, f, \mathrm{d}) \le h_K\|(I-P_{2k-2}^{(2)})(f)\|_{L^2(K)} + c(\|(I - P_{k-1}^{(\infty)})(\nabla\cdot\mathrm{d})\|_{\boldsymbol{L}^\infty(K)} + \|(I - P_k^{(\infty)})(\mathrm{d})\|_{\mathbb{L}^\infty(K)})\|\nabla u_h\|_{\boldsymbol{L}^2(K)}$ with $(\nabla\cdot\mathrm{d})_i := \sum_{j\in\{1:d\}} \frac{\partial}{\partial x_j}\mathrm{d}_{ji}$ for all $i \in \{1{:}d\}$. Prove that $\phi_F^{\mathrm{s}}(u_h, f, \mathrm{d}) \le c\|(I - P_k^{(\infty)})(\mathrm{d})\|_{\mathbb{L}^\infty(F)}\|\nabla u_h\|_{\boldsymbol{L}^2(D_F)}$ with best-approximation operator $P_k^{(\infty)}$ mapping to $L^\infty(F)$. What are the decay rates of the oscillation terms for smooth f and d? (iii) What happens if $l^{\mathrm{v}} := k$ and $l^{\mathrm{s}} := k-1$ for piecewise constant d?

Exercise 34.5 (Error reduction). Consider two discrete spaces $V_{h_1} \subset V_{h_2} \subset H_0^1(D)$ with corresponding discrete solutions u_{h_1} and u_{h_2}, respectively. Consider the norm $\|v\|_a := a(v,v)^{\frac12}$ for all $v \in H_0^1(D)$. Prove that $\|u - u_{h_1}\|_a^2 = \|u - u_{h_2}\|_a^2 + \|u_{h_2} - u_{h_1}\|_a^2$. (*Hint*: use the Galerkin orthogonality property.)

Exercise 34.6 (Approximation class for smooth solution). Let D be a Lipschitz polyhedron in $\mathbb{R}^d$. Prove that $H^{k+1}(D) \subset A_{k/d}$. (*Hint*: consider uniformly refined meshes.)

Exercise 34.7 (Graded mesh). Let $D := (0,1)$ and let $(x_i)_{i\in\{0:I\}}$, $I \geq 2$, be a mesh of D. Let $u \in W^{1,1}(D)$ and consider the piecewise constant function u_I such that $u_I(x) := u(x_{i-1})$ for all $x \in (x_{i-1}, x_i)$ and all $i \in \{1{:}I\}$. (i) Assume $u \in W^{1,\infty}(D)$. Prove that the decay rate $\|u - u_I\|_{L^\infty(D)} \leq \frac{1}{I}\|u'\|_{L^\infty(D)}$ is achieved using a uniform mesh. (ii) Assume now $u \in W^{1,1}(D)$. Prove that the decay rate $\|u - u_I\|_{L^\infty(D)} \leq \frac{1}{I}\|u'\|_{L^1(D)}$ is achieved using a graded mesh such that $x_i := \Phi^{(-1)}(\frac{i}{I})$, where $\Phi(s) := \frac{1}{\|u'\|_{L^1(D)}}\int_0^s |u'(t)|\,\mathrm{d}t$ for all $s \in (0,1)$ and all $i \in \{0{:}I\}$.

Solution to exercises

Exercise 34.1 (Residual). We observe that

$$\begin{aligned}\langle \rho(u_h), \varphi\rangle &= \sum_{K\in\mathcal{T}_h}\int_K (f\varphi - (\mathfrak{d}\nabla u_h)\cdot\nabla\varphi)\,\mathrm{d}x\\ &= \sum_{K\in\mathcal{T}_h}\int_K (f + \nabla\cdot(\mathfrak{d}\nabla u_h))\varphi\,\mathrm{d}x - \sum_{K\in\mathcal{T}_h}\int_{\partial K}((\mathfrak{d}\nabla u_h)\cdot\boldsymbol{n}_K)\varphi\,\mathrm{d}s,\end{aligned}$$

where $\boldsymbol{n}_K$ denotes the outward unit normal to ∂K. We conclude by regrouping the terms from both sides of each interface and observing that φ vanishes at the boundary faces.

Exercise 34.2 (Trace inequality). Let $\boldsymbol{z} \in \mathcal{V}_h$ and let $v \in H^1_*(D_{\boldsymbol{z}})$. Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}^\circ_{\boldsymbol{z}}$. Proceeding as in the proof of Lemma 12.15 with $p := 2$, we infer that

$$\frac{|K|}{|F|}\|v\|^2_{L^2(F)} \leq \|v\|^2_{L^2(K)} + 2d^{-1}h_K\|v\|_{L^2(K)}\|\nabla v\|_{\boldsymbol{L}^2(K)},$$

where $K \in \{K_l, K_r\}$ is one of the two cells sharing F. Let $D_F := \mathrm{int}(K_l \cup K_r)$. Summing over these two cells, using $h_K \leq h_{D_{\boldsymbol{z}}}$, and the Cauchy–Schwarz inequality for the rightmost term yielding $\sum_{K\in\{K_l,K_r\}}\|v\|_{L^2(K)}\|\nabla v\|_{\boldsymbol{L}^2(K)} \leq \|v\|_{L^2(D_F)}\|\nabla v\|_{\boldsymbol{L}^2(D_F)}$, we arrive at

$$\frac{|D_F|}{|F|}\|v\|^2_{L^2(F)} \leq \|v\|^2_{L^2(D_F)} + 2d^{-1}h_{D_{\boldsymbol{z}}}\|v\|_{L^2(D_F)}\|\nabla v\|_{\boldsymbol{L}^2(D_F)}.$$

We now sum over all the faces $F \in \mathcal{F}^\circ_{\boldsymbol{z}}$. Since any mesh cell in $\mathcal{T}_{\boldsymbol{z}}$ has exactly d faces sharing the vertex $\boldsymbol{z}$, we have

$$\sum_{F\in\mathcal{F}^\circ_{\boldsymbol{z}}}\|v\|^2_{L^2(D_F)} = d\|v\|^2_{L^2(D_{\boldsymbol{z}})}.$$

Invoking the Cauchy–Schwarz inequality yields

$$\begin{aligned}\sum_{F\in\mathcal{F}^\circ_{\boldsymbol{z}}}\|v\|_{L^2(D_F)}\|\nabla v\|_{\boldsymbol{L}^2(D_F)} &\leq \Big(\sum_{F\in\mathcal{F}^\circ_{\boldsymbol{z}}}\|v\|^2_{L^2(D_F)}\Big)^{\frac12}\Big(\sum_{F\in\mathcal{F}^\circ_{\boldsymbol{z}}}\|\nabla v\|^2_{\boldsymbol{L}^2(D_F)}\Big)^{\frac12}\\ &\leq d\|v\|_{L^2(D_{\boldsymbol{z}})}\|\nabla v\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}.\end{aligned}$$

We infer that

$$\sum_{F\in\mathcal{F}^\circ_{\boldsymbol{z}}}\frac{|D_F|}{|F|}\|v\|^2_{L^2(F)} \leq d\|v\|^2_{L^2(D_{\boldsymbol{z}})} + 2h_{D_{\boldsymbol{z}}}\|v\|_{L^2(D_{\boldsymbol{z}})}\|\nabla v\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}.$$

Finally, the definition of $\varrho_{\boldsymbol{z}}$ implies that

$$\varrho_{\boldsymbol{z}}^{-1} h_{D_{\boldsymbol{z}}} \|v\|_{L^2(\mathcal{F}_{\boldsymbol{z}}^\circ)}^2 \le d\|v\|_{L^2(D_{\boldsymbol{z}})}^2 + 2h_{D_{\boldsymbol{z}}}\|v\|_{L^2(D_{\boldsymbol{z}})}\|\nabla v\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})},$$

where we used that $\|v\|_{L^2(\mathcal{F}_{\boldsymbol{z}}^\circ)}^2 := \sum_{F\in\mathcal{F}_{\boldsymbol{z}}^\circ}\|v\|_{L^2(F)}^2$. We conclude by invoking Definition 34.5 which implies that

$$\varrho_{\boldsymbol{z}}^{-1} h_{D_{\boldsymbol{z}}}^{-1} \|v\|_{L^2(\mathcal{F}_{\boldsymbol{z}}^\circ)}^2 \le (dC_{\mathrm{PS},\boldsymbol{z}}^2 + 2C_{\mathrm{PS},\boldsymbol{z}})\|\nabla v\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}^2.$$

Exercise 34.3 (Bound on dual norm). (i) Invoking the Poincaré–Steklov inequality on the reference simplex and transferring back to K by pullback implies that $\|\varphi\|_{L^2(K)} \le ch_K\|\nabla\varphi\|_{\boldsymbol{L}^2(K)}$ for all $\varphi\in H_0^1(K)$. Using this inequality and the Cauchy–Schwarz inequality, we infer that

$$\|T_K^{\mathrm{v}}(f)\|_{H^{-1}(K)} = \sup_{\varphi\in H_0^1(K)} \frac{|\int_K f\varphi\,\mathrm{d}x|}{\|\nabla\varphi\|_{\boldsymbol{L}^2(K)}} \le c\,h_K\|f\|_{L^2(K)}.$$

(ii) Let $F := \partial K_l\cap\partial K_r \in \mathcal{F}_h^\circ$. Let $\varphi\in H_0^1(D_F)$. The Poincaré–Steklov inequality (proved as above on the reference simplex and transferred by pullback) yields $\|\varphi\|_{L^2(K)} \le ch_K\|\nabla\varphi\|_{\boldsymbol{L}^2(K)}$ for all $K\in\mathcal{T}_F = \{K_l, K_r\}$. Combining this bound with the multiplicative trace inequality from Lemma 12.15 and using the regularity of the mesh sequence, we infer that $\|\varphi\|_{L^2(F)} \le ch_F^{\frac12}\|\nabla\varphi\|_{\boldsymbol{L}^2(K)}$. We can now conclude as above.

Exercise 34.4 (Oscillation). (i) Let $v\in\mathbb{P}_n$. Since $(P_{m-n}^{(\infty)}\theta)v_h\in\mathbb{P}_m$, we observe that

$$\begin{aligned}\|\theta v_h - P_m^{(2)}(\theta v_h)\|_{L^2(K)} &\le \|\theta v_h - (P_{m-n}^{(\infty)}\theta)v_h\|_{L^2(K)}\\ &= \|(\theta - P_{m-n}^{(\infty)}\theta)v_h\|_{L^2(K)} \le \|\theta - P_{m-n}^{(\infty)}\theta\|_{L^\infty(K)}\|v_h\|_{L^2(K)}.\end{aligned}$$

(ii) Since $f - \nabla\cdot(\mathfrak{d}\nabla u_h) = f - (\nabla\cdot\mathfrak{d})\cdot\nabla u_h - \mathfrak{d}:D^2u_h$, where D^2u_h denotes the Hessian matrix of u_h, we infer using the triangle inequality that

$$\begin{aligned}\phi_K^{\mathrm{v}}(u_h,f,\mathfrak{d}) &= h_K\|(I-P_{2k-2}^{(2)})(f-\nabla\cdot(\mathfrak{d}\nabla u_h))\|_{L^2(K)}\\ &= h_K\|(I-P_{2k-2}^{(2)})f\|_{L^2(K)} + h_K\|(I-P_{2k-2}^{(2)})((\nabla\cdot\mathfrak{d})\cdot\nabla u_h)\|_{L^2(K)}\\ &\quad + h_K\|(I-P_{2k-2}^{(2)})(\mathfrak{d}:D^2u_h)\|_{L^2(K)}.\end{aligned}$$

We conclude using the result from Step (i) componentwise for the last two terms on the right-hand side together with an inverse inequality on the Hessian of u_h. To prove the bound on $\phi_F^{\mathrm{s}}(u_h,f,\mathfrak{d})$, we first observe that $[\![\mathfrak{d}\nabla u_h]\!]_F\cdot\boldsymbol{n}_F = (\mathfrak{d}\nabla u_h)_{|K_l}\cdot\boldsymbol{n}_{K_lF} + (\mathfrak{d}\nabla u_h)_{|K_r}\cdot\boldsymbol{n}_{K_rF}$. Using the triangle inequality and best-approximation operators in $L^p(F)$, we obtain

$$\phi_F^{\mathrm{s}}(u_h,f,\mathfrak{d}) \le h_F^{\frac12}\|(I-P_{2k-1}^{(2)})(\mathfrak{d}\nabla u_h)_{|K_l}\|_{L^2(F)} + h_F^{\frac12}\|(I-P_{2k-1}^{(2)})(\mathfrak{d}\nabla u_h)_{|K_r}\|_{L^2(F)}.$$

Finally, we use the result from Step (i) together with a discrete trace inequality and the regularity of the mesh sequence. If $f_{|K}$ and $\mathfrak{d}_{|K}$ are smooth, namely $f_{|K}\in H^{k-1}(K)$ and $\mathfrak{d}_{|K}\in W^{k,\infty}(K;\mathbb{R}^{d\times d})$, we infer that

$$\begin{aligned}\phi_K^{\mathrm{v}}(u_h,f,\mathfrak{d}) &\le ch_K^{k+1}(\|f\|_{H^{k+1}(K)} + \|\nabla u_h\|_{\boldsymbol{L}^2(K)}),\\ \phi_F^{\mathrm{s}}(u_h,f,\mathfrak{d}) &\le ch_F^{k+1}\|\nabla u_h\|_{\boldsymbol{L}^2(D_F)}.\end{aligned}$$

(iii) If $\mathfrak{d}$ is piecewise constant, choosing $l^{\mathrm{v}} := k$ and $l^{\mathrm{s}} := k-1$ leads to $\phi_K^{\mathrm{v}} = h_K\|f-\bar f\|_{L^2(K)}$ and $\phi_K^{\mathrm{s}} = 0$, where $\bar f$ is the L^2-orthogonal projection of f onto $P_k(K)$. This implies that ϕ_K^{v} superconverges by two orders with respect to the approximation error.

Exercise 34.5 (Error reduction). Since $u-u_{h_1}=(u-u_{h_2})+(u_{h_2}-u_{h_1})$, the conclusion follows from

$$\|u-u_{h_1}\|_a^2=\|u-u_{h_2}\|_a^2+\|u_{h_2}-u_{h_1}\|_a^2+2a(u-u_{h_2},u_{h_2}-u_{h_1}),$$

owing to the symmetry of a, and the last term on the right-hand side vanishes owing to the Galerkin orthogonality property since $u_{h_2}-u_{h_1}\in V_{h_2}$.

Exercise 34.6 (Approximation class for smooth solution). Let D be a Lipschitz polyhedron in $\mathbb{R}^d$. Let $u\in H^{k+1}(D)$. Let $(\mathcal{T}_n)_{n\in\mathbb{N}}$ be a quasi-uniform sequence of matching affine meshes (see Definition 22.20) so that each mesh $\mathcal{T}_h$ covers exactly D. Let h_n denote the maximal diameter of the cells composing $\mathcal{T}_n$. The quasi-uniformity of the sequence implies that the d-dimensional measure of every mesh cell is uniformly equivalent to h_n^d, i.e., there is c s.t. $\operatorname{card}(\mathcal{T}_n)\le ch_n^{-d}|D|$. Moreover, we have established in Corollary 22.9 that

$$\inf_{v_h\in P_k^g(\mathcal{T}_n)}\|\nabla(u-v)\|_{\boldsymbol{L}^2(D)}\le c\,h_n^k|u|_{H^{k+1}(D)}.$$

Hence, we have

$$\inf_{v_h\in P_k^g(\mathcal{T}_n)}\|\nabla(u-v)\|_{\boldsymbol{L}^2(D)}\le c\,\operatorname{card}(\mathcal{T}_n)^{-\frac{k}{d}}|D|^{\frac{k}{d}}|u|_{H^{k+1}(D)}.$$

This implies that

$$|u|_{A_{\frac{k}{d}}}\le c\,|D|^{\frac{k}{d}}|u|_{H^{k+1}(D)},$$

i.e., $u\in A_{\frac{k}{d}}$. This proves that $H^{k+1}(D)\subset A_{\frac{k}{d}}$.

Exercise 34.7 (Graded mesh). (i) Let $x\in D$. There is $i\in\{1{:}I\}$ such that $x\in(x_{i-1},x_i)$. We infer that

$$|u(x)-u_I(x)|=|u(x)-u(x_{i-1})|\le\int_{x_{i-1}}^x|u'(t)|\,\mathrm{d}t\le|x_i-x_{i-1}|\,\|u'\|_{L^\infty(D)}.$$

This proves the assertion on a uniform mesh since we have $|x_i-x_{i-1}|=\frac{1}{I}$.
(ii) We first observe that

$$\frac{1}{\|u'\|_{L^1(D)}}\int_{x_{i-1}}^{x_i}|u'(t)|\,\mathrm{d}t=\Phi(x_i)-\Phi(x_{i-1})=\frac{1}{I}.$$

As a result, we infer that

$$|u(x)-u_I(x)|=|u(x)-u(x_{i-1})|\le\int_{x_{i-1}}^{x_i}|u'(t)|\,\mathrm{d}t=\frac{1}{I}\|u'\|_{L^1(D)}.$$

Chapter 35

The Helmholtz problem

Exercises

Exercise 35.1 (1D Helmholtz, well-posedness). Let $D := (0, \ell_D)$, $\kappa > 0$, and consider the Helmholtz problem with mixed boundary conditions: $-\partial_{xx} u - \kappa^2 u = f$ in D, $u(0) = 0$, and $\partial_x u(\ell_D) - \mathrm{i}\kappa u(\ell_D) = 0$. (i) Give a weak formulation in $V := \{v \in H^1(D) \mid v(0) = 0\}$. (ii) Show by invoking an ODE argument that if the weak formulation has a solution, then it is unique. (iii) Show that the weak problem is well-posed. (*Hint*: use Lemma 35.3.)

Exercise 35.2 (Green's function, 1D). Let $G : D \times D \to \mathbb{C}$ be the function defined by

$$G(x,s) := \kappa^{-1} \begin{cases} \sin(\kappa x) e^{\mathrm{i}\kappa s} & \text{if } x \in [0,s], \\ \sin(\kappa s) e^{\mathrm{i}\kappa x} & \text{if } x \in [s,1]. \end{cases}$$

(i) Prove that for all $x \in D$, the function $D \ni s \mapsto G(x,s) \in \mathbb{C}$ solves the PDE $-\partial_{ss} u - \kappa^2 u = \delta_{s=x}$ in D with the boundary conditions $u(0) = 0$ and $\partial_s u(\ell_D) - \mathrm{i}\kappa u(\ell_D) = 0$ (i.e., G is the Green's function of the Helmholtz problem from Exercise 35.1). (ii) Find $H(x,s)$ s.t. $\partial_s H(x,s) = \partial_x G(x,s)$. (iii) Let $u(x) := \int_0^{\ell_D} G(x,s) f(s) \,\mathrm{d}s$. Prove that $\|u\|_{L^2(D)} \le \kappa^{-1} \|f\|_{L^2(D)}$, $|u|_{H^1(D)} \le \|f\|_{L^2(D)}$, and $|u|_{H^2(D)} \le (\kappa+1)\|f\|_{L^2(D)}$. (iv) Let $v \in L^2(D)$ and let $\widetilde{z}(x) := \kappa^2 \int_0^{\ell_D} G(x,s) v(s) \,\mathrm{d}s$. What is the PDE solved by $\widetilde{z}$? Same question for $z(x) := \kappa^2 \int_0^{\ell_D} \overline{G}(x,s) v(s) \,\mathrm{d}s$. *Note*: The function z is invoked in Step (1) of the proof of Theorem 35.11. (v) Assume now that $v \in H^1(D)$ with $v(0) = 0$, and let z and $\widetilde{z}$ be defined as above. Prove that $\max(|z|_{H^1(D)}, |\widetilde{z}|_{H^1(D)}) \le 4\kappa\ell_D |v|_{H^1(D)}$. (*Hint*: see Ihlenburg and Babuška [29, p. 14] (up to the factor 4).)

Exercise 35.3 (Variation on Fortin's lemma). Let V, W be two Banach spaces and let a be a bounded sesquilinear form on $V \times W$ like in Fortin's Lemma 26.9. Let $(V_h)_{h\in\mathcal{H}}$, $(W_h)_{h\in\mathcal{H}}$ be sequences of subspaces of V and W equipped with the norm of V and W, respectively. Assume that there exists a map $\Pi_h : W \to W_h$ and constants $\gamma_{\Pi_h} > 0$, $c(h) > 0$ such that $|a(v_h, w - \Pi_h(w))| \le c(h)\|v_h\|_V \|w\|_W$, $\gamma_{\Pi_h} \|\Pi_h(w)\|_W \le \|w\|_W$ for all $v_h \in V_h$, all $w \in W$, and all $h \in \mathcal{H}$. Assume that $\lim_{h\to 0} c(h) = 0$. Prove that the discrete inf-sup condition (26.5a) holds true for $h \in \mathcal{H}$ small enough.

Exercise 35.4 (Lemma 35.8). (i) Prove that $\Re((\boldsymbol{m}\cdot\nabla v)\overline{v}) = \frac{1}{2}\boldsymbol{m}\cdot\nabla|v|^2$ for all $v \in H^1(D;\mathbb{C})$ and $\boldsymbol{m} \in \mathbb{R}^d$. (ii) Prove that $\Re(\boldsymbol{m}\cdot((\nabla \boldsymbol{v})^{\mathsf{T}}\overline{\boldsymbol{v}})) = \frac{1}{2}\boldsymbol{m}\cdot\nabla\|\boldsymbol{v}\|^2_{\ell^2(\mathbb{C}^d)}$ for all $\boldsymbol{v} \in H^1(D;\mathbb{C}^d)$ and $\boldsymbol{m} \in \mathbb{R}^d$.

(iii) Let $q \in H^2(D;\mathbb{C})$ and let D^2q denote the Hessian matrix of q, i.e., $(D^2q)_{ij} = \partial^2_{x_ix_j}q$ for all $i,j \in \{1{:}d\}$. Show that $\Re(\boldsymbol{m}\cdot((D^2q)\nabla\overline{q})) = \frac{1}{2}\boldsymbol{m}\cdot\nabla\|\nabla q\|^2_{\ell^2(\mathbb{C}^d)}$. (iv) Prove that (35.11) holds true for all $q \in \{v \in H^1(D;\mathbb{C}) \mid \Delta v \in L^2(D;\mathbb{C}),\ \nabla v \in L^2(\partial D;\mathbb{C}^d)\}$ and all $\boldsymbol{m} \in W^{1,\infty}(D;\mathbb{R}^d)$. (*Hint*: assume first that $q \in H^2(D;\mathbb{C})$.)

Solution to exercises

Exercise 35.1 (1D Helmholtz, well-posedness). (i) One possible weak formulation is as follows: Find $u \in V$ such that for all $v \in V$,

$$\int_0^{\ell_D} (\partial_x u \partial_x v - \kappa^2 uv)\,\mathrm{d}x - \mathsf{i}\kappa u(\ell_D)v(\ell_D) = \int_0^{\ell_D} f(x)v(x)\,\mathrm{d}x.$$

(ii) Let us consider the homogeneous problem and let u be a solution to the homogeneous problem $a(u,w) = 0$ for all $w \in V$. This implies that $0 = |a(u,u)| \geq \kappa u(\ell_D)^2$. Hence, $u(\ell_D) = 0$, and the Robin condition implies that $\partial_x u(\ell_D) = 0$ as well. In conclusion, we have

$$\partial_{xx}u - \kappa^2 u = 0, \quad \partial_x u(\ell_D) = 0, \quad u(\ell_D) = 0.$$

This is a linear second-order ODE with homogeneous data. The unique solution is $u = 0$.
(iii) The well-posedness follows by invoking Lemma 35.3 since the bilinear form $\int_0^{\ell_D}(\partial_x u\partial_x v - \kappa^2 uv)\,\mathrm{d}x - \mathsf{i}\kappa u(\ell_D)v(\ell_D)$ satisfies the inequality (35.4a).

Exercise 35.2 (Green's function, 1D). (i) Let $x \in D$ be fixed. We observe that

$$\begin{aligned} &G(x,0) = 0,\\ &\partial_s G(x,\ell_D) - \mathsf{i}\kappa G(x,\ell_D) = \mathsf{i}\sin(\kappa\ell_D)e^{\mathsf{i}\kappa\ell_D} - \mathsf{i}\sin(\kappa\ell_D)e^{\mathsf{i}\kappa\ell_D} = 0. \end{aligned}$$

Moreover, it is clear that $G(x,s)$ is continuous at x. We now have to verify that

$$-\partial_{ss}G(x,s) - \kappa^2 G(x,s) = \delta_{x=s},$$

where $\delta_{x=s}$ is the Dirac measure whose support is $\{x\}$. We observe that

$$-\partial_{ss}G(x,s) - \kappa^2 G(x,s) = \begin{cases} \kappa\sin(\kappa x)e^{\mathsf{i}\kappa s} - \kappa\sin(\kappa x)e^{\mathsf{i}\kappa s} = 0 & \text{if } x \in [0,s],\\ \kappa\sin(\kappa s)e^{\mathsf{i}\kappa x} - \kappa\sin(\kappa s)e^{\mathsf{i}\kappa x} = 0 & \text{if } x \in [s,1]. \end{cases}$$

Let us now verify the jump condition $-\big(\partial_s G(x,x^+) - \partial_s G(x,x^-)\big) = 1$, which, let us recall, together with the above identity and the continuity of $G(x,\cdot)$, is equivalent to stating that $\langle -\partial_{ss}G(x,\cdot) - \kappa^2 G(x,\cdot),\varphi\rangle = \varphi(x)$ for all $\varphi \in C_0^\infty(D)$. We indeed have

$$-\big(\partial_s G(x,x^+) - \partial_s G(x,x^-)\big) = -\mathsf{i}\sin(\kappa x)e^{\mathsf{i}kx} + \cos(\kappa x)e^{\mathsf{i}kx} = e^{-\mathsf{i}kx}e^{\mathsf{i}kx} = 1.$$

(ii) Let $H(x,s) := \int_0^s \partial_x G(x,t)\,\mathrm{d}t$. We first consider the case $s \leq x$. This yields

$$H(x,s) = \kappa^{-1}\int_0^s \mathsf{i}\kappa\sin(\kappa t)e^{\mathsf{i}\kappa x}\,\mathrm{d}t = -\mathsf{i}\kappa^{-1}(\cos(\kappa s) - 1)e^{\mathsf{i}\kappa x}.$$

In the second case $x \le s$, we have

$$\begin{aligned} H(x,s) &= \kappa^{-1}\int_0^x \mathrm{i}\kappa\sin(\kappa t)e^{\mathrm{i}\kappa x}\,\mathrm{d}t + \kappa^{-1}\int_x^s \kappa\cos(\kappa x)e^{\mathrm{i}\kappa t}\,\mathrm{d}t \\ &= -\mathrm{i}\kappa^{-1}(\cos(\kappa x)-1)e^{\mathrm{i}\kappa x} - \mathrm{i}\kappa^{-1}\cos(\kappa x)(e^{\mathrm{i}\kappa s}-1) \\ &= \mathrm{i}\kappa^{-1}(e^{\mathrm{i}\kappa x} - \cos(\kappa x)e^{\mathrm{i}\kappa s}) = -\mathrm{i}\kappa^{-1}e^{\mathrm{i}\kappa x}(\cos(\kappa x)e^{\mathrm{i}\kappa(s-x)}-1). \end{aligned}$$

Notice that in both cases, we have $|H(x,s)| \le 2\kappa^{-1}$.
(iii) We have

$$|u(x)| = \left|\int_0^{\ell_D} G(x,s)f(s)\,\mathrm{d}s\right| \le \ell_D^{\frac12}\|G(x,\cdot)\|_{L^\infty(D)}\|f\|_{L^2(D)}.$$

Hence, $\|u\|_{L^2(D)} \le \ell_D\kappa^{-1}\|f\|_{L^2(D)}$ because $\|G(x,\cdot)\|_{L^\infty(D)} \le \kappa^{-1}$. Moreover, we have

$$|\partial_x u(x)| = \left|\int_0^{\ell_D} \partial_x G(x,s)f(s)\,\mathrm{d}s\right| \le \|\partial_x G(x,\cdot)\|_{L^\infty(D)}\ell_D^{\frac12}\|f\|_{L^2(D)}.$$

This implies that $|u|_{H^1(D)} \le \ell_D\|f\|_{L^2(D)}$. Recall that since G is the Green's function of the Helmholtz problem from Exercise 35.1, we have $\partial_{xx}u - \kappa^2 u = f$, $u(0)=0$, and $\partial_x u(\ell_D) - \mathrm{i}\kappa u(\ell_D) = 0$. Hence, we can estimate $|u|_{H^2(D)}$ as follows:

$$|u|_{H^2(D)} = \|\kappa^2 u + f\|_{L^2(D)} \le \kappa^2\|u\|_{L^2(D)} + \|f\|_{L^2(D)} \le (\kappa+1)\|f\|_{L^2(D)}.$$

(iv) Let us assume that $v \in H^1(D)$ and $v(0)=0$. Let us set $\widetilde{z}(x) := \kappa^2\int_0^{\ell_D} G(x,s)v(s)\,\mathrm{d}s$. Since G is the Green's function of the Helmholtz problem from Exercise 35.1, $\widetilde{z}$ solves

$$\partial_{xx}\widetilde{z} - \kappa^2\widetilde{z} = \kappa^2 v, \quad \widetilde{z}(0) = 0, \quad \partial_x\widetilde{z}(\ell_D) - \mathrm{i}\kappa\widetilde{z}(\ell_D) = 0.$$

Let us now set $z(x) := \kappa^2\int_0^{\ell_D}\overline{G}(x,s)v(s)\,\mathrm{d}s$. Since $\overline{G}$ is the Green's function of the adjoint problem, z solves

$$\partial_{xx}z - \kappa^2 z = \kappa^2 v, \quad z(0) = 0, \quad \partial_x z(\ell_D) + \mathrm{i}\kappa z(\ell_D) = 0.$$

(iv) Using that $v(\ell_D) = \int_0^{\ell_D}\partial_s v(s)\,\mathrm{d}s$, we infer that $|v(\ell_D)| \le \ell_D^{\frac12}|v|_{H^1(D)}$, which in turn implies that

$$\begin{aligned} \kappa^{-2}\partial_x z(x) &= \int_0^{\ell_D}\partial_x G(x,s)v(s)\,\mathrm{d}s = \int_0^{\ell_D}\partial_s H(x,s)v(s)\,\mathrm{d}s \\ &= -\int_0^{\ell_D} H(x,s)\partial_s v(s)\,\mathrm{d}s + H(x,\ell_D)v(\ell_D) \\ &\le \|H(x,\cdot)\|_{L^\infty(D)}\big(\ell_D^{\frac12}|v|_{H^1(D)} + |v(\ell_D)|\big) \\ &\le 2\|H(x,\cdot)\|_{L^\infty(D)}\ell_D^{\frac12}|v|_{H^1(D)}. \end{aligned}$$

Hence, $|z|_{H^1(D)} \le 4\kappa\ell_D|v|_{H^1(D)}$ because $\|H(x,\cdot)\|_{L^\infty(D)} \le 2\kappa^{-1}$. The same argument holds true for $\widetilde{z}$.

Exercise 35.3 (Variation on Fortin's lemma). Let $v_h \in V_h$. Using the assumptions, we have

$$\begin{aligned}\sup_{w_h\in W_h}\frac{|a(v_h,w_h)|}{\|w_h\|_W} &\ge \sup_{w\in W}\frac{|a(v_h,\Pi_h(w))|}{\|\Pi_h(w)\|_W} \ge \gamma_{\Pi_h}\sup_{w\in W}\frac{|a(v_h,\Pi_h(w))|}{\|w\|_W}\\ &\ge \gamma_{\Pi_h}\sup_{w\in W}\frac{|a(v_h,w)|}{\|w\|_W} - \gamma_{\Pi_h}\sup_{w\in W}\frac{|a(v_h,\Pi_h(w)-w)|}{\|w\|_W}\\ &\ge \gamma_{\Pi_h}\alpha\|v_h\|_V - \gamma_{\Pi_h}c(h)\|v_h\|_V.\end{aligned}$$

Let ℓ_0 be such that $c(h)\le\frac12\alpha$ for all $h\in(0,\ell_0]$. We have

$$\inf_{v_h\in V_h}\sup_{w_h\in W_h}\frac{|a(v_h,w_h)|}{\|w_h\|_W} \ge \frac12\gamma_{\Pi_h}\alpha,$$

for all $h\in(0,\ell_0]$. This proves (26.5a) with $\alpha_h\ge\frac12\gamma_{\Pi_h}\alpha$ for all $h\in(0,\ell_0]$.

Exercise 35.4 (Lemma 35.8). (i) We have

$$\begin{aligned}2\Re((\boldsymbol{m}\cdot\nabla v)\overline{v}) &= (\boldsymbol{m}\cdot\nabla v)\overline{v} + (\boldsymbol{m}\cdot\nabla\overline{v})v\\ &= \boldsymbol{m}\cdot\nabla(v\overline{v}) - (\boldsymbol{m}\cdot\nabla\overline{v})v + (\boldsymbol{m}\cdot\nabla\overline{v})v = \boldsymbol{m}\cdot\nabla|v|^2,\end{aligned}$$

which proves the result.
(ii) Recalling that $(\nabla\boldsymbol{v})_{ij}=\partial_{x_j}v_i$ for all $i,j\in\{1{:}d\}$, we can apply the above identity as follows:

$$\Re(\boldsymbol{m}\cdot((\nabla\boldsymbol{v})^{\mathsf{T}}\overline{\boldsymbol{v}})) = \sum_{j\in\{1:d\}}\Re((\boldsymbol{m}\cdot\nabla v_j)\overline{v}_j) = \frac12\sum_{j\in\{1:d\}}\boldsymbol{m}\cdot\nabla|v_j|^2 = \frac12\boldsymbol{m}\cdot\nabla\|\boldsymbol{v}\|^2_{\ell^2(\mathbb{C}^d)},$$

which proves the result.
(iii) Let $q\in H^2(D;\mathbb{C})$. Using $\boldsymbol{v}=\nabla q$ in the identity from Step (ii) and recalling that D^2q is a symmetric matrix leads to $\Re(\boldsymbol{m}\cdot(D^2q\nabla\overline{q})) = \frac12\boldsymbol{m}\cdot\nabla\|\nabla q\|^2_{\ell^2(\mathbb{C}^d)}$.
(iv) Assume first that $q\in H^2(D;\mathbb{C})$ and let $\boldsymbol{m}\in W^{1,\infty}(D;\mathbb{R}^d)$. Integration by parts gives

$$\begin{aligned}-\int_D\Delta q(\boldsymbol{m}\cdot\nabla\overline{q})\,\mathrm{d}x &= \int_D\nabla q\cdot\nabla(\boldsymbol{m}\cdot\nabla\overline{q})\,\mathrm{d}x - \int_{\partial D}(\boldsymbol{n}\cdot\nabla q)(\boldsymbol{m}\cdot\nabla\overline{q})\,\mathrm{d}s\\ &= \int_D\nabla q\cdot((\nabla\boldsymbol{m})^{\mathsf{T}}\nabla\overline{q})\,\mathrm{d}x + \int_D\boldsymbol{m}\cdot((D^2\overline{q})\nabla q)\,\mathrm{d}x - \int_{\partial D}(\boldsymbol{n}\cdot\nabla q)(\boldsymbol{m}\cdot\nabla\overline{q})\,\mathrm{d}s.\end{aligned}$$

We now apply the identity established in Step (iii) integrated over D. Integrating by parts leads to

$$\begin{aligned}\Re\left(\int_D\boldsymbol{m}\cdot((D^2\overline{q})\nabla q)\,\mathrm{d}x\right) &= \int_D\frac12\boldsymbol{m}\cdot\nabla\|\nabla q\|^2_{\ell^2(\mathbb{C}^d)}\,\mathrm{d}x\\ &= -\int_D\frac12(\nabla\cdot\boldsymbol{m})\|\nabla q\|^2_{\ell^2(\mathbb{C}^d)}\,\mathrm{d}x + \int_{\partial D}(\boldsymbol{m}\cdot\boldsymbol{n})\|\nabla q\|^2_{\ell^2(\mathbb{C}^d)}\,\mathrm{d}s.\end{aligned}$$

Notice that all the integrations by parts make sense since q and $\boldsymbol{m}$ have sufficient smoothness. Putting everything together, we infer that the identity (35.11) holds true for all $q\in H^2(D;\mathbb{C})$ and all $\boldsymbol{m}\in W^{1,\infty}(D;\mathbb{R}^d)$. Reasoning as in the second step of the proof of Lemma 35.7, i.e., invoking a density argument, we conclude that this identity still holds true if $q\in\{v\in H^1(D;\mathbb{C})\mid\Delta v\in L^2(D;\mathbb{C}),\ \nabla v\in L^2(\partial D;\mathbb{C}^d)\}$.

Chapter 36

Crouzeix–Raviart approximation

Exercises

Exercise 36.1 (Commuting properties). Let K be a simplex in $\mathbb{R}^d$ and let Π_K^0 denote the L^2-orthogonal projection onto constants. Prove that $\nabla(\mathcal{I}_K^{\mathrm{CR}}(p)) = \Pi_K^0(\nabla p)$ and $\nabla\cdot(\boldsymbol{\mathcal{I}}_K^{\mathrm{CR}}(\boldsymbol{\sigma})) = \Pi_K^0(\nabla\cdot\boldsymbol{\sigma})$ for all $p \in H^1(K)$ and all $\boldsymbol{\sigma} \in \boldsymbol{L}^2(K)$ with $\nabla\cdot\boldsymbol{\sigma} \in L^1(K)$ and $\boldsymbol{\mathcal{I}}_K^{\mathrm{CR}}$ defined componentwise using $\mathcal{I}_h^{\mathrm{CR}}$.

Exercise 36.2 (Best approximation). Let $v \in H^1(D)$. A global best-approximation of v in $P_1^{\mathrm{CR}}(\mathcal{T}_h)$ in the broken H^1-seminorm is a function $v_h^{\mathrm{CR}} \in P_1^{\mathrm{CR}}(\mathcal{T}_h)$ s.t.

$$\sum_{K\in\mathcal{T}_h} \|\nabla(v - v_h^{\mathrm{CR}})\|_{\boldsymbol{L}^2(K)}^2 = \min_{v_h\in P_1^{\mathrm{CR}}(\mathcal{T}_h)} \sum_{K\in\mathcal{T}_h} \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(K)}^2.$$

(i) Write a characterization of v_h^{CR} in weak form and show that v_h^{CR} is unique up to an additive constant. (*Hint*: adapt Proposition 25.8.) (ii) Let v_h^{b} be a global best-approximation of v in the broken finite element space $P_1^{\mathrm{b}}(\mathcal{T}_h)$; see §32.2. Prove that $\sum_{K\in\mathcal{T}_h} \|\nabla(v - v_h^{\mathrm{CR}})\|_{\boldsymbol{L}^2(K)}^2 = \sum_{K\in\mathcal{T}_h} \|\nabla(v - v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(K)}^2$. (*Hint*: using Exercise 36.1, show that $v_h^{\mathrm{CR}} = \mathcal{I}_h^{\mathrm{CR}}(v)$ up to an additive constant.)

Exercise 36.3 ($\boldsymbol{H}$(div)-flux recovery). Let u_h solve (36.10). Assume that f is piecewise constant on $\mathcal{T}_h$. Set $\boldsymbol{\sigma}_{h|K} := -\nabla u_{h|K} + \frac{1}{d} f_{|K}(\boldsymbol{x} - \boldsymbol{x}_K)$, where $\boldsymbol{x}_K$ is the barycenter of K for all $K \in \mathcal{T}_h$. Prove that $\boldsymbol{\sigma}_h$ is in the lowest-order Raviart–Thomas finite element space $\boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$ and that $\nabla\cdot\boldsymbol{\sigma} = f$; see Marini [33] (*Hint*: evaluate $\int_F [\![\boldsymbol{\sigma}_h]\!]\cdot\boldsymbol{n}_F \varphi_F^{\mathrm{CR}} \,\mathrm{d}s$ for all $F \in \mathcal{F}_h^\circ$.)

Exercise 36.4 (Discrete Helmholtz). Let $D \subset \mathbb{R}^2$ be a simply connected polygon. Prove that $\boldsymbol{P}_0^{\mathrm{b}}(\mathcal{T}_h) = \nabla P_1^{\mathrm{g}}(\mathcal{T}_h) \oplus \nabla_h^\perp P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$, where

$$\nabla_h^\perp P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h) := \{\boldsymbol{v}_h \in \boldsymbol{P}_0^{\mathrm{b}}(\mathcal{T}_h) \mid \exists q_h \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h) \mid \boldsymbol{v}_{h|K} = \nabla^\perp(q_{h|K}), \forall K \in \mathcal{T}_h\},$$

and $\nabla^\perp$ is the two-dimensional curl operator defined in Remark 16.17. (*Hint*: prove that the decomposition is L^2-orthogonal and use a dimension argument based on Euler's relations.)

Exercise 36.5 (Rannacher–Turek). Let $K := [-1, 1]^d$. For all $i \in \{1{:}d\}$ and $\alpha \in \{l, r\}$, let $F_{i,\alpha}$ be the face of K corresponding to $\{x_i = -1\}$ when $\alpha = l$ and to $\{x_i = 1\}$ when $\alpha = r$.

Observe that there are $2d$ such faces, each of measure 2^{d-1}. Let P be spanned by the $2d$ functions $\{1, x_1, \dots, x_d, x_1^2 - x_2^2, \dots, x_{d-1}^2 - x_d^2\}$. Consider the linear forms $\sigma_{i,\alpha}(p) := 2^{1-d} \int_{F_{i,\alpha}} p \,\mathrm{d}s$ for all $i \in \{1{:}d\}$ and $\alpha \in \{l, r\}$. Setting $\Sigma := \{\sigma_{i,\alpha}\}_{i\in\{1:d\},\alpha\in\{l,r\}}$, prove that (K, P, Σ) is a finite element. *Note*: this element has been introduced by [40] for the mixed discretization of the Stokes equations on Cartesian grids.

Exercise 36.6 (Quadratic space). Let $\mathcal{T}_h$ be a triangulation of a simply connected domain $D \subset \mathbb{R}^2$ and let

$$P_2^{\mathrm{CR}}(\mathcal{T}_h) := \{v_h \in P_2^{\mathrm{b}}(\mathcal{T}_h) \mid \int_F [\![v_h]\!]_F (q \circ \boldsymbol{T}_F^{-1}) \,\mathrm{d}s = 0,\ \forall F \in \mathcal{F}_h^\circ,\ \forall q \in \mathbb{P}_{1,1}\},$$

where $\boldsymbol{T}_F$ is an affine bijective mapping from the unit segment $\widehat{S}^1 = [-1, 1]$ to F. Orient all the faces $F \in \mathcal{F}_h$ and define the two Gauss points $\boldsymbol{g}_F^\pm$ on F that are the image by $\boldsymbol{T}_F$ of $\widehat{g}^\pm := \pm\frac{\sqrt{3}}{3}$, in such a way that the orientation of F goes from $\boldsymbol{g}_F^-$ to $\boldsymbol{g}_F^+$. For all $K \in \mathcal{T}_h$, let $\{\lambda_{0,K}, \lambda_{1,K}, \lambda_{2,K}\}$ be the barycentric coordinates in K and set $b_K := 2 - 3(\lambda_{0,K}^2 + \lambda_{1,K}^2 + \lambda_{2,K}^2)$ (this function is usually called Fortin–Soulié bubble [17]). One can verify that a polynomial $p \in \mathbb{P}_{2,2}$ vanishes at the six points $\{\boldsymbol{g}_F^\pm\}_{F\in\mathcal{F}_K}$ if and only if $p = \alpha b_K$ for some $\alpha \in \mathbb{R}$. *Note*: this shows that these six points, which lie on an ellipse, cannot be taken as nodes of a $\mathbb{P}_{2,2}$ Lagrange element. (i) Extending b_K by zero outside K, verify that $b_K \in P_2^{\mathrm{CR}}(\mathcal{T}_h)$. (ii) Set $B := \mathrm{span}_{K\in\mathcal{T}_h}\{b_K\}$ and $B_* := \{v_h \in B \mid \int_D v_h \,\mathrm{d}x = 0\}$. Prove that $P_2^{\mathrm{g}}(\mathcal{T}_h) + B_* \subset P_2^{\mathrm{CR}}(\mathcal{T}_h)$ and that $P_2^{\mathrm{g}}(\mathcal{T}_h) \cap B_* = \{0\}$. (iii) Define $J : P_2^{\mathrm{CR}}(\mathcal{T}_h) \to \mathbb{R}^{2N_{\mathrm{f}}}$ s.t. $J(v_h) := (v_h(\boldsymbol{g}_F^-), v_h(\boldsymbol{g}_F^+))_{F\in\mathcal{F}_h}$ for all $v_h \in P_2^{\mathrm{CR}}(\mathcal{T}_h)$. Prove that $\dim(\ker(J)) = N_{\mathrm{c}}$ and $\dim(\mathrm{im}(J)) \le 2N_{\mathrm{f}} - N_{\mathrm{c}}$. (*Hint*: any polynomial $p \in \mathbb{P}_{2,2}$ satisfies $\sum_{F\in\mathcal{F}_K}(p(\boldsymbol{g}_F^+) - p(\boldsymbol{g}_F^-)) = 0$ for all $K \in \mathcal{T}_h$.) (iv) Prove that $P_2^{\mathrm{CR}}(\mathcal{T}_h) = P_2^{\mathrm{g}}(\mathcal{T}_h) \oplus B_*$; see Greff [19]. (*Hint*: use a dimensional argument and Euler's relation from Remark 8.13.)

Solution to exercises

Exercise 36.1 (Commuting properties). Let $p \in H^1(K)$. We observe that

$$\int_K \nabla p \,\mathrm{d}x = \sum_{F\in\mathcal{F}_K} \int_F p\boldsymbol{n}_K \,\mathrm{d}s = \sum_{F\in\mathcal{F}_K} \int_F \mathcal{I}_K^{\mathrm{CR}}(p)\boldsymbol{n}_K \,\mathrm{d}s = \int_K \nabla(\mathcal{I}_h^{\mathrm{CR}}(p)) \,\mathrm{d}x.$$

Since $\nabla(\mathcal{I}_h^{\mathrm{CR}}(p))$ is constant on K, we conclude that $\nabla(\mathcal{I}_K^{\mathrm{CR}}(p)) = \Pi_K^0(\nabla p)$. Let $\boldsymbol{\sigma} \in \boldsymbol{L}^2(K)$ with $\nabla{\cdot}\boldsymbol{\sigma} \in L^1(K)$. We observe that

$$\int_K \nabla{\cdot}\boldsymbol{\sigma} \,\mathrm{d}x = \sum_{F\in\mathcal{F}_K} \int_F \boldsymbol{\sigma}{\cdot}\boldsymbol{n}_K \,\mathrm{d}s = \sum_{F\in\mathcal{F}_K} \int_F \boldsymbol{\mathcal{I}}_K^{\mathrm{CR}}(\boldsymbol{\sigma}){\cdot}\boldsymbol{n}_K \,\mathrm{d}s = \int_K \nabla{\cdot}(\boldsymbol{\mathcal{I}}_h^{\mathrm{CR}}(\boldsymbol{\sigma})) \,\mathrm{d}x.$$

Since $\nabla{\cdot}(\boldsymbol{\mathcal{I}}_h^{\mathrm{CR}}(\boldsymbol{\sigma}))$ is constant on K, we conclude that $\nabla{\cdot}(\boldsymbol{\mathcal{I}}_K^{\mathrm{CR}}(\boldsymbol{\sigma})) = \Pi_K^0(\nabla{\cdot}\boldsymbol{\sigma})$.

Exercise 36.2 (Best approximation). (i) The function v_h^{CR} is a minimizer in $P_1^{\mathrm{CR}}(\mathcal{T}_h)$ of the functional

$$\tilde{\mathfrak{E}}(w_h) = \frac{1}{2} \sum_{K\in\mathcal{T}_h} (\nabla(w_h - v), \nabla(w_h - v))_{\boldsymbol{L}^2(K)},$$

or, equivalently (since the function $v \in H^1(D)$ is fixed), of the functional

$$\mathfrak{E}(w_h) = \frac{1}{2} \sum_{K\in\mathcal{T}_h} (\nabla w_h, \nabla w_h)_{\boldsymbol{L}^2(K)} - \sum_{K\in\mathcal{T}_h} (\nabla v, \nabla w_h)_{\boldsymbol{L}^2(K)}.$$

Reasoning as in the proof of Proposition 25.8, we infer that v_h^{CR} is s.t.

$$\sum_{K\in\mathcal{T}_h}(\nabla v_h^{\mathrm{CR}},\nabla w_h)_{\boldsymbol{L}^2(K)}=\sum_{K\in\mathcal{T}_h}(\nabla v,\nabla w_h)_{\boldsymbol{L}^2(K)},\qquad\forall w_h\in P_1^{\mathrm{CR}}(\mathcal{T}_h).$$

The bilinear form $\sum_{K\in\mathcal{T}_h}(\nabla w_h,\nabla w_h)_{\boldsymbol{L}^2(K)}$ is coercive on the subspace $\{w_h\in P_1^{\mathrm{CR}}(\mathcal{T}_h)\mid\int_D w_h\,\mathrm{d}x=0\}$. This shows the existence of v_h^{CR} and its uniqueness up to an additive constant.
(ii) For all $w_h\in P_k^{\mathrm{b}}(\mathcal{T}_h)$, we have

$$\sum_{K\in\mathcal{T}_h}(\nabla(\mathcal{I}_h^{\mathrm{CR}}(v)),\nabla w_h)_{\boldsymbol{L}^2(K)}=\sum_{K\in\mathcal{T}_h}(\Pi_K^0(\nabla v),\nabla w_h)_{\boldsymbol{L}^2(K)}=\sum_{K\in\mathcal{T}_h}(\nabla v,\nabla w_h)_{\boldsymbol{L}^2(K)},$$

where the last equality follows from the fact that ∇w_h is piecewise constant. This shows that $\mathcal{I}_h^{\mathrm{CR}}(v)=v_h^{\mathrm{b}}$, up to an additive constant in each mesh cell, and restricting the test function to $w_h\in P_1^{\mathrm{CR}}(\mathcal{T}_h)$, we infer that $\mathcal{I}_h^{\mathrm{CR}}(v)=v_h^{\mathrm{CR}}$ up to a global additive constant. Therefore, we have the expected identity

$$\sum_{K\in\mathcal{T}_h}\|\nabla(v-v_h^{\mathrm{CR}})\|_{\boldsymbol{L}^2(K)}^2=\sum_{K\in\mathcal{T}_h}\|\nabla(v-v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(K)}^2.$$

Exercise 36.3 ($\boldsymbol{H}$(div)-flux recovery). By definition $\boldsymbol{\sigma}_{h|K}\in\mathbb{RT}_{0,d}$ and $\nabla\cdot(\boldsymbol{\sigma}_{h|K})=f_{|K}$ for all $K\in\mathcal{T}_h$. Let $F\in\mathcal{F}_h^\circ$ with $F:=\partial K_l\cap\partial K_r$. We infer that

$$\begin{aligned}\int_F[\![\boldsymbol{\sigma}_h]\!]\cdot\boldsymbol{n}_F\varphi_F^{\mathrm{CR}}\,\mathrm{d}s&=\int_{K_l\cup K_r}(\nabla\cdot\boldsymbol{\sigma}_h)\varphi_F^{\mathrm{CR}}\,\mathrm{d}x+\int_{K_l\cup K_r}\boldsymbol{\sigma}_h\cdot\nabla\varphi_F^{\mathrm{CR}}\,\mathrm{d}x\\&=\int_{K_l\cup K_r}f\varphi_F^{\mathrm{CR}}\,\mathrm{d}x-\int_{K_l\cup K_r}\nabla u_h\cdot\nabla\varphi_F^{\mathrm{CR}}\,\mathrm{d}x=0,\end{aligned}$$

since $\int_{K_l\cup K_r}(\boldsymbol{x}-\boldsymbol{x}_K)\cdot\nabla\varphi_F^{\mathrm{CR}}\,\mathrm{d}x=0$ and $\mathrm{supp}(\varphi_F^{\mathrm{CR}})=\mathrm{int}(K_l\cup K_r)$. This implies that $[\![\boldsymbol{\sigma}_h]\!]\cdot\boldsymbol{n}_F=0$ for all $F\in\mathcal{F}_h^\circ$ since the normal component of a function in $\mathbb{RT}_{0,d}$ is constant on each face; see Lemma 14.7. We conclude that $\boldsymbol{\sigma}_h\in\boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$ since the zero-jump condition implies that $\boldsymbol{\sigma}_h\in\boldsymbol{H}(\mathrm{div};D)$ owing to Theorem 18.10. Since $\boldsymbol{\sigma}_h$ is in $\boldsymbol{H}(\mathrm{div};D)$, its divergence equals its piecewise divergence, i.e., $\nabla\cdot\boldsymbol{\sigma}_h=f$ in D.

Exercise 36.4 (Discrete Helmholtz). Let $p_h\in P_1^{\mathrm{g}}(\mathcal{T}_h)$ and let $q_h\in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. Integrating by parts cellwise, we infer that

$$(\nabla p_h,\nabla_h^\perp q_h)_{L^2(D)}=\sum_{K\in\mathcal{T}_h}\int_{\partial K}(\nabla^\perp p_h\cdot\boldsymbol{n}_K)q_h\,\mathrm{d}s.$$

Since $\nabla^\perp p_h\cdot\boldsymbol{n}_K$ is constant on each face of K and since q_h has continuous mean value on all the mesh interfaces and zero mean value on all the boundary faces, we conclude that

$$(\nabla p_h,\nabla_h^\perp q_h)_{L^2(D)}=\sum_{F\in\mathcal{F}_h^\circ}\int_F[\![\nabla^\perp p_h]\!]\cdot\boldsymbol{n}_Fq_h\,\mathrm{d}s=0,$$

since $\nabla^\perp p_{h|K_l}\cdot\boldsymbol{n}_{K_l}=-\nabla^\perp p_{h|K_r}\cdot\boldsymbol{n}_{K_r}$ for all $F:=\partial K_l\cap\partial K_r$ owing to the continuity of p_h across F (note that $\nabla^\perp p_{h|K}\cdot\boldsymbol{n}_K$ only depends on the tangential derivatives of p_h on ∂K). Moreover, we have $\dim(\nabla P_1^{\mathrm{g}}(\mathcal{T}_h))=N_{\mathrm{v}}-1$ and $\dim(\nabla_h^\perp P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h))=N_{\mathrm{e}}-N_{\mathrm{e}}^\partial$ since $\dim(P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h))=N_{\mathrm{e}}-N_{\mathrm{e}}^\partial$ and $\nabla_h^\perp$ is injective on $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. Using Euler's relations from Remark 8.13, we obtain

$$\dim(\nabla P_1^{\mathrm{g}}(\mathcal{T}_h))+\dim(\nabla_h^\perp P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h))=N_{\mathrm{v}}-1+N_{\mathrm{e}}-N_{\mathrm{e}}^\partial=2N_{\mathrm{c}}=\dim(\boldsymbol{P}_0^{\mathrm{b}}(\mathcal{T}_h)).$$

We conclude that $\boldsymbol{P}_0^{\mathrm{b}}(\mathcal{T}_h)=\nabla P_1^{\mathrm{g}}(\mathcal{T}_h)\oplus\nabla_h^\perp P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$.

Exercise 36.5 (Rannacher–Turek). The $2d$ functions

$$(1, x_1, \dots, x_d, x_1^2 - x_2^2, \dots, x_{d-1}^2 - x_d^2)$$

are linearly independent. Hence,

$$\operatorname{card}\Sigma = \dim P = 2d.$$

Consider the linear combination $\sum_{\alpha\in\{l,r\}}\sum_{i\in\{1:d\}}\beta_{i,\alpha}\sigma_{i,\alpha}$ and assume that it is the zero form in $\mathcal{L}(P;\mathbb{R})$. Consider first $p := x_j$ for $j \in \{1{:}d\}$. Then $\sigma_{i,\alpha}(p) = 0$ if $j \neq i$, whereas $\sigma_{j,l}(p) = -1$ and $\sigma_{j,r}(p) = 1$. Hence, $\beta_{i,l} = \beta_{i,r}$ for all $i \in \{1{:}d\}$. Consider then $p := x_j^2 - x_{j+1}^2$ for $j \in \{1{:}(d-1)\}$, so that $\sigma_{i,\alpha}(p) = 0$ if $i \notin \{j, j+1\}$, whereas $\sigma_{j,l}(p) = \sigma_{j,r}(p) = \frac{1}{3}$ and $\sigma_{j+1,l}(p) = \sigma_{j+1,r}(p) = -\frac{1}{3}$. Hence, $\beta_{j,l} = \beta_{j+1,l}$ for all $j \in \{1{:}(d-1)\}$. As a consequence, all the coefficients $\beta_{i,\alpha}$ take the same value, say β, and considering $p := 1$ for which $\sigma_{i,\alpha}(p) = 1$ for all $i \in \{1{:}d\}$ and $\alpha \in \{l, r\}$, we infer that $\beta(2d) = 0$, whence $\beta = 0$.

Exercise 36.6 (Quadratic space). (i) For all $K \in \mathcal{T}_h$ and all $F \in \mathcal{F}_K$, b_K vanishes at the points $\{\boldsymbol{g}_F^{\pm}\}$. Let indeed $\lambda_{i,K}, \lambda_{j,K}$ be the two barycentric coordinates that do not vanish on F. Thus, $(\lambda_{i,K}^2 \circ \boldsymbol{T}_F)(\widehat{g}^{\pm}) + (\lambda_{j,K}^2 \circ \boldsymbol{T}_F)(\widehat{g}^{\pm}) = \frac{2}{3}$. This shows that $(b_{K|F} \circ \boldsymbol{T}_F)(\widehat{g}^{\pm}) = 0$. Hence, $\int_F b_K(q \circ \boldsymbol{T}_F^{-1})\,\mathrm{d}s = \int_{\widehat{F}}(b_K \circ \boldsymbol{T}_F)q\,\mathrm{d}\widehat{s} = 0$ for all $q \in \mathbb{P}_{1,1}$, since $(b_K \circ \boldsymbol{T}_F)q$ is a polynomial of degree three and the two-point quadrature based on $\widehat{g}^{\pm}$ is exact for polynomials of degree three. Since $[\![b_K]\!]_F = \pm b_K$, we infer that $\int_F [\![b_K]\!]_F(q \circ \boldsymbol{T}_F^{-1})\,\mathrm{d}s = 0$. Since $b_K \in P_2^{\mathrm{b}}(\mathcal{T}_h)$, we obtain $b_K \in P_2^{\mathrm{CR}}(\mathcal{T}_h)$.
(ii) Since $P_2^{\mathrm{g}}(\mathcal{T}_h) \subset P_2^{\mathrm{CR}}(\mathcal{T}_h)$ and $B_* \subset B \subset P_2^{\mathrm{CR}}(\mathcal{T}_h)$, we infer that $P_2^{\mathrm{g}}(\mathcal{T}_h) + B_* \subset P_2^{\mathrm{CR}}(\mathcal{T}_h)$. Let now $v_h \in P_2^{\mathrm{g}}(\mathcal{T}_h) \cap B_*$ so that $v_h := \sum_{K\in\mathcal{T}_h}\alpha_K b_K$. Let $F \in \mathcal{F}_h^{\circ}$ and let K_l, K_r be the two cells such that $F := \partial K_l \cap \partial K_r$. Let $\lambda_{i_l,K_l}, \lambda_{j_l,K_l}$ and $\lambda_{i_r,K_r}, \lambda_{j_r,K_r}$ be the barycentric coordinates in K_l and K_r, respectively, that are nonzero over F. The continuity of v_h across F implies that

$$\alpha_{K_l}(2 - 3\lambda_{i_l,K_l}^2 - 3\lambda_{j_l,K_l}^2)_{|F} = \alpha_{K_r}(2 - 3\lambda_{i_r,K_r}^2 - 3\lambda_{j_r,K_r}^2)_{|F},$$

which implies that $\alpha_{K_l} = \alpha_{K_r}$ since $(\lambda_{i_l,K_l}^2 + \lambda_{j_l,K_l}^2)_{|F} = (\lambda_{i_r,K_r}^2 + \lambda_{j_r,K_r}^2)_{|F}$. Hence, $v_h = \alpha\sum_{K\in\mathcal{T}_h} b_K$. Moreover, a direct computation gives $\int_K b_K\,\mathrm{d}x = \frac{1}{2}|K|$. Hence, $\int_D v_h\,\mathrm{d}x = \frac{1}{2}\alpha|D|$. Finally, $v_h \in B_*$ implies that $\int_D v_h\,\mathrm{d}x = 0$, so that $\alpha = 0$.
(iii) A function v_h is in $\ker(J)$ if it vanishes at the six points $\{\boldsymbol{g}_F^{\pm}\}_{F\in\mathcal{F}_K}$ for all $K \in \mathcal{T}_h$. Hence, $\ker(J) = B$ so that $\dim(\ker(J)) = N_{\mathrm{c}}$. Let us now consider $\mathrm{im}(J)$. For all $K \in \mathcal{T}_h$, consider the vector $\psi_K \in \mathbb{R}^{2N_{\mathrm{f}}}$ with components $(\psi_{K,F^{\pm}})_{F\in\mathcal{F}_h}$ such that $\psi_{K,F^-} = -1$ and $\psi_{K,F^+} = 1$ if $F \in \mathcal{F}_K$, and $\psi_{K,F^{\pm}} = 0$ otherwise. Then, the hint means that any vector in $\mathrm{im}(J)$ is orthogonal (for the Euclidean inner product) to ψ_K for all $K \in \mathcal{T}_h$. It remains to show that the family $\{\psi_K\}_{K\in\mathcal{T}_h}$ is linearly independent. Assume that $\sum_{K\in\mathcal{T}_h}\mu_K\psi_K = 0$ in $\mathbb{R}^{2N_{\mathrm{f}}}$. Considering the two components of this vector attached to an interface $F := \partial K_l \cap \partial K_r$, we infer that $\mu_{K_l} = \mu_{K_r}$. Hence, $\mu_K := \mu_0$ for all $K \in \mathcal{T}_h$. Finally, considering a boundary face $F := \partial K_l \cap \partial D$, we obtain $\mu_0 = \mu_{K_l} = 0$.
(iv) We observe that

$$\begin{aligned}
2N_{\mathrm{f}} &= N_{\mathrm{f}} + (N_{\mathrm{v}} + N_{\mathrm{c}} - 1) = (N_{\mathrm{v}} + N_{\mathrm{f}}) + (N_{\mathrm{c}} - 1)\\
&= \dim(P_2^{\mathrm{g}}(\mathcal{T}_h)) + \dim(B_*) = \dim(P_2^{\mathrm{g}}(\mathcal{T}_h) \oplus B_*)\\
&\le \dim(P_2^{\mathrm{CR}}(\mathcal{T}_h)) = \dim(\ker(J)) + \dim(\mathrm{im}(J))\\
&\le N_{\mathrm{c}} + 2N_{\mathrm{f}} - N_{\mathrm{c}} = 2N_{\mathrm{f}},
\end{aligned}$$

where we used Euler's relation (see Remark 8.13), the inclusion $P_2^{\mathrm{g}}(\mathcal{T}_h) \oplus B_* \subset P_2^{\mathrm{CR}}(\mathcal{T}_h)$ from Step (ii), and the rank nullity theorem for J together with Step (iii). Hence, the above inequalities are equalities, showing that $\dim(P_2^{\mathrm{g}}(\mathcal{T}_h) \oplus B_*) = \dim(P_2^{\mathrm{CR}}(\mathcal{T}_h))$. Since $P_2^{\mathrm{g}}(\mathcal{T}_h) \oplus B_* \subset P_2^{\mathrm{CR}}(\mathcal{T}_h)$, we conclude that the reverse conclusion also holds true.

Chapter 37

Nitsche's boundary penalty method

Exercises

Exercise 37.1 (Poincaré–Steklov). Let $\check{C}_{\mathrm{PS}}$ be defined in (31.23). Prove that $\check{C}_{\mathrm{PS}}\ell_D^{-1}\|v\|_{L^2(D)} \leq (\|\nabla v\|^2_{\boldsymbol{L}^2(D)} + |v|^2_\partial)^{\frac{1}{2}}$ for all $v \in H^1(D)$. (*Hint*: use $h \leq \ell_D$ and (31.23).)

Exercise 37.2 (Quadratic inequality). Prove that $x^2 - 2\beta xy + \varpi_0 y^2 \geq \frac{\varpi_0 - \beta^2}{1+\varpi_0}(x^2+y^2)$ for all real numbers x, y, $\varpi_0 \geq 0$ and $\beta \geq 0$.

Exercise 37.3 (Error estimate). Prove (37.14). (*Hint*: consider the quasi-interpolation operator from §22.3.)

Exercise 37.4 (Gradient). Let U be an open bounded set in $\mathbb{R}^d$, let $s \in (0,1)$, and set $\boldsymbol{H}^s_{00}(U) := [\boldsymbol{L}^2(U), \boldsymbol{H}^1_0(U)]_{s,2}$. (i) Show that $\nabla : \boldsymbol{H}^{1-s}(U) \to (\boldsymbol{H}^s_{00}(U))'$ is bounded for all $s \in (0,1)$. (*Hint*: use Theorems A.27 and A.30.) (ii) Assume that U is Lipschitz. Show that $\nabla : \boldsymbol{H}^{1-s}(U) \to \boldsymbol{H}^{-s}(U)$ is bounded for all $s \in (0,1)$, $s \neq \frac{1}{2}$. (*Hint*: see (3.7), Theorem 3.19; see also Grisvard [20, Lem. 1.4.4.6].)

Exercise 37.5 (L^2-estimate). (i) Modify the proof of Theorem 37.7 by measuring the interpolation error on the adjoint solution with the operator $\mathcal{I}_h^{\mathrm{g,av}}$ instead of $\mathcal{I}_{h0}^{\mathrm{g,av}}$, i.e., use $Y_h := V_h$ instead of $Y_h := V_h \cap H^1_0(D)$. (*Hint*: set $a_\sharp(v,w) := (\nabla v, \nabla w)_{\boldsymbol{L}^2(D)} - (\boldsymbol{n}\cdot\nabla v, w)_{L^2(\partial D)} + \sum_{F\in\mathcal{F}_h^\partial} \varpi_0 \frac{1}{h_F}(v,w)_{L^2(F)}$.) (ii) Do the same for the proof of Theorem 37.8.

Solution to exercises

Exercise 37.1 (Poincaré–Steklov). Since $h \leq \ell_D$, we have

$$\|\nabla v\|^2_{\boldsymbol{L}^2(D)} + |v|^2_\partial \geq \|\nabla v\|^2_{\boldsymbol{L}^2(D)} + \ell_D^{-1}\|v\|^2_{L^2(\partial D)},$$

where $\ell_D := \mathrm{diam}(D)$. Using the Poincaré–Steklov inequality (31.23) leads to the expected bound.

Exercise 37.2 (Quadratic inequality). Notice that $x^2 - 2\beta xy + \varpi_0 y^2 \geq \frac{\varpi_0-\beta^2}{1+\varpi_0}(x^2+y^2)$ iff

$$\frac{1+\beta^2}{1+\varpi_0}x^2 - 2\beta xy + \frac{\varpi_0^2+\beta^2}{1+\varpi_0}y^2 \geq 0.$$

Since the coefficients $\frac{1+\beta^2}{1+\varpi_0}$ and $\frac{\varpi_0^2+\beta^2}{1+\varpi_0}$ are both positive, the above quadratic form is nonnegative iff

$$0 \leq \frac{1+\beta^2}{1+\varpi_0}\frac{\varpi_0^2+\beta^2}{1+\varpi_0} - \beta^2.$$

Rearranging the terms leads to $2\varpi_0\beta^2 \leq \varpi_0^2 + \beta^4$, which is trivially true.

Exercise 37.3 (Error estimate). We bound the infimum in (37.13) by taking $v_h := \mathcal{I}_h^{\mathrm{g,av}}(u)$, where $\mathcal{I}_h^{\mathrm{g,av}} : L^1(D) \to P_k^{\mathrm{g}}(\mathcal{T}_h)$ is the quasi-interpolation operator from §22.3. Let us localize the $\|\cdot\|_{V_\sharp}$-norm as follows:

$$\|v\|_{V_\sharp}^2 := \sum_{K\in\mathcal{T}_h} \|v\|_{V_\sharp(K)}^2,$$
$$\|v\|_{V_\sharp(K)}^2 := \|\nabla v\|_{\boldsymbol{L}^2(K)}^2 + h_F^{-1}\|v\|_{L^2(F)}^2 + \sum_{F\in\mathcal{F}_K\cap\mathcal{F}_h^\partial} h_F\|\boldsymbol{n}\cdot\nabla v\|_{L^2(F)}^2,$$

if $K \in \mathcal{T}_h^{\partial D}$ and $\|v\|_{V_\sharp(K)}^2 = \|\nabla v\|_{\boldsymbol{L}^2(K)}^2$ otherwise. Owing to the estimate (22.14) from Theorem 22.6 (with $m := 1$, $p := 2$) and to the approximation results for $\mathcal{I}_h^{\mathrm{g,av}}$ on faces (see Exercise 22.5), we infer that $\|u - \mathcal{I}_h^{\mathrm{g,av}}(u)\|_{V_\sharp(K)} \leq ch_K^t|u|_{H^{1+t}(\check{\mathcal{T}}_K)}$ for all $K \in \mathcal{T}_h$, where $\check{\mathcal{T}}_K$ is the collection of the mesh cells sharing at least a vertex with K. Then (37.14) follows by invoking the regularity of the mesh sequence.

Exercise 37.4 (Gradient). (i) Let $y \in L^2(U)$. By definition of the weak derivative of y, we have

$$\|\nabla y\|_{\boldsymbol{H}^{-1}(U)} = \sup_{\boldsymbol{v}\in\boldsymbol{H}_0^1(U)} \frac{|\langle\nabla y, \boldsymbol{v}\rangle|}{\|\boldsymbol{v}\|_{\boldsymbol{H}_0^1(U)}} = \sup_{\boldsymbol{v}\in\boldsymbol{H}_0^1(U)} \frac{|\int_U y\nabla\cdot\boldsymbol{v}\,\mathrm{d}x|}{\|\boldsymbol{v}\|_{\boldsymbol{H}_0^1(U)}} \leq c\,\|y\|_{L^2(U)}.$$

Let now $y \in H^1(U)$. We then have

$$\|\nabla y\|_{\boldsymbol{L}^2(U)} \leq c\,\|y\|_{H^1(U)}.$$

This shows that ∇ maps boundedly from $L^2(U)$ to $\boldsymbol{H}^{-1}(U)$ and from $H^1(U)$ to $\boldsymbol{L}^2(U)$. The Riesz–Thorin theorem (Theorem A.27) implies that ∇ maps boundedly from $[L^2(U), H^1(U)]_{1-s,2} = H^{1-s}(U)$ to $[\boldsymbol{H}^{-1}(U), \boldsymbol{L}^2(U)]_{1-s,2}$ for all $s \in (0,1)$. But Theorem A.30 implies that

$$[\boldsymbol{H}^{-1}(U), \boldsymbol{L}^2(U)]_{1-s,2} = [\boldsymbol{L}^2(U), \boldsymbol{H}_0^1(U)]'_{s,2}.$$

Setting $\boldsymbol{H}_{00}^s(U) := [\boldsymbol{L}^2(U), \boldsymbol{H}_0^1(U)]_{s,2}$, we have

$$[\boldsymbol{H}^{-1}(U), \boldsymbol{L}^2(U)]_{1-s,2} = (\boldsymbol{H}_{00}^s(U))'.$$

(ii) From (3.7) and Theorem 3.19, we know that $\boldsymbol{H}_{00}^s(U) = \boldsymbol{H}_0^s(U)$ for all $s \in (0,1)$ if $s \neq \frac{1}{2}$. Hence, ∇ maps boundedly from $H^{1-s}(U)$ to $\boldsymbol{H}^{-s}(U) := (\boldsymbol{H}_0^s(U))'$ for all $s \in (0,1)$ if $s \neq \frac{1}{2}$.

Exercise 37.5 (L^2-estimate). (i) Let us consider $V_\sharp := V_{\mathrm{s}} + V_h$, $Z_{\mathrm{s}} := H^{1+s}(D) \cap H_0^1(D)$, $Y_h := V_h$, and $Z_\sharp := Z_{\mathrm{s}} + V_h$ equipped with the norm $\|z\|_{Z_\sharp}^2 := \|\nabla z\|_{\boldsymbol{L}^2(D)}^2 + |z|_\partial^2$. Let us consider the bilinear form

$$a_\sharp(v,w) := (\nabla v, \nabla w)_{\boldsymbol{L}^2(D)} - (\boldsymbol{n}\cdot\nabla v, w)_{L^2(\partial D)} + \sum_{F\in\mathcal{F}_h^\partial} \varpi_0 h_F^{-1}(v,w)_{L^2(F)}.$$

Notice that $a_\sharp$ is bounded on $V_\sharp \times Z_\sharp$. The Galerkin orthogonality property holds true for $a_\sharp$ since for all $y_h \in V_h$,

$$a_\sharp(u, y_h) = \ell_h(y_h) = a_h(u_h, y_h) = a_\sharp(u_h, y_h).$$

Let $\delta^{\mathrm{adj}}(\zeta_e)$ be the adjoint consistency error defined in (36.29), i.e., for all $v \in V_\sharp$,

$$\langle \delta^{\mathrm{adj}}(\zeta_e), v\rangle_{V_\sharp', V_\sharp} := -(v, \Delta\zeta_e)_{L^2(D)} - a_\sharp(v, \zeta_e).$$

Since ζ_e vanishes on ∂D, the following identity holds true: For all $v \in V_\sharp$,

$$\langle \delta^{\mathrm{adj}}(\zeta_e), v\rangle_{V_\sharp', V_\sharp} = -(v, \boldsymbol{n}\cdot\nabla\zeta_e)_{L^2(\partial D)}.$$

Hence, the adjoint consistency error can be bounded as in the proof of Theorem 37.7. Concerning the interpolation error on the adjoint solution, we can now consider the interpolation operator $\mathcal{I}_h^{\mathrm{g,av}}$ from §22.3, and we deduce as before that

$$\inf_{y_h \in Y_h} \|\nabla(\zeta_e - y_h)\|_{\boldsymbol{L}^2(D)} \le \|\zeta_e - \mathcal{I}_h^{\mathrm{g,av}}(\zeta_e)\|_{Z_\sharp} \le c\, h^s |\zeta_e|_{H^{1+s}(D)}.$$

(ii) For the proof of Theorem 37.8, one proceeds as above by considering the bilinear form $a_\sharp(v, w) := (\nabla v, \nabla w)_{\boldsymbol{L}^2(D)} - (\boldsymbol{n}\cdot\nabla v, w)_{L^2(\partial D)} - (v, \boldsymbol{n}\cdot\nabla w)_{L^2(\partial D)} + \sum_{F\in\mathcal{F}_h^\partial} \varpi_0 h_F^{-1}(v, w)_{L^2(F)}$.

Chapter 38

Discontinuous Galerkin

Exercises

Exercise 38.1 (Elementary dG identities). (i) Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$. Prove that $2\{\boldsymbol{\sigma}\cdot\boldsymbol{n}_K q\} = (\{\boldsymbol{\sigma}\}[\![q]\!] + [\![\boldsymbol{\sigma}]\!]\{q\})\cdot\boldsymbol{n}_F$. (ii) Let $\theta_l, \theta_r \in [0,1]$ such that $\theta_l + \theta_r = 1$. Let $[\![a]\!]_\theta := 2(\theta_r a_l - \theta_l a_r)$ and $\{a\}_\theta := \theta_l a_l + \theta_r a_r$. Show that $\{ab\} = \{a\}\{b\}_\theta + \frac{1}{4}[\![a]\!]_\theta[\![b]\!]$.

Exercise 38.2 (Boundary conditions). (i) Assume that u solves the Poisson problem (38.1) with the non-homogeneous Dirichlet condition $u = g$ on ∂D. Let a_h^θ be defined in (38.20). Devise $\ell_h^{\theta,\mathrm{nD}}$ so that exact consistency holds for the following formulation: Find $u_h \in V_h$ such that $a_h^\theta(u_h, w_h) = \ell_h^{\theta,\mathrm{nD}}(w_h)$ for all $w_h \in V_h$. (ii) Assume that u solves the Poisson problem with the Robin condition $\gamma u + \boldsymbol{n}\cdot\nabla u = g$ on ∂D. Let ℓ_h^{Rb} be defined in (38.13b). Devise a_h^{Rb} so that exact consistency holds for the following formulation: Find $u_h \in V_h$ such that $a_h^{\theta,\mathrm{Rb}}(u_h, w_h) = \ell_h^{\mathrm{Rb}}(w_h)$ for all $w_h \in V_h$.

Exercise 38.3 (L^2-estimate). Prove Theorem 38.12. (*Hint*: see the proof of Theorem 37.8.)

Exercise 38.4 (Local lifting). Prove (38.22a). (*Hint*: use (38.10).)

Exercise 38.5 (Local formulation). Write the local formulation of the OBB, NIP, and IIP dG methods discussed in Remark 38.13.

Exercise 38.6 (Extending (38.25)). Let $\tilde{a}_h$ (resp., a_h) be defined by extending (38.25) (resp., (38.4)) to $V_\sharp \times V_h$. Show that $\tilde{a}_h(v, w_h) = a_h(v, w_h) + \sum_{F\in\mathcal{F}_h}\int_F\{\nabla_h v - \mathcal{I}_h^{\mathrm{b}}(\nabla_h v)\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s$ for all $(v, w_h) \in V_\sharp \times V_h$.

Exercise 38.7 (Discrete gradient). Let $(v_h)_{h\in\mathcal{H}}$ be a sequence in $(V_h)_{h\in\mathcal{H}}$ (meaning that $v_h \in V_h$ for all $h \in \mathcal{H}$). Assume that there is C s.t. $\|v_h\|_{V_h} \le C$ for all $h \in \mathcal{H}$. One can show that there is $v \in L^2(D)$ such that, up to a subsequence, $v_h \to v$ in $L^2(D)$ as $h \to 0$; see [15, Thm. 5.6]. (i) Show that, up to a subsequence, $\mathfrak{G}_h^l(v_h)$ weakly converges to some $\boldsymbol{G}$ in $\boldsymbol{L}^2(D)$ as $h \to 0$. (*Hint*: bound $\|\mathfrak{G}_h^l(v_h)\|_{\boldsymbol{L}^2(D)}$.) (ii) Show that $\boldsymbol{G} = \nabla v$ and that $v \in H_0^1(D)$. (*Hint*: extend functions by zero outside D and prove first that $\int_{\mathbb{R}^d}\mathfrak{G}_h^l(v_h)\cdot\boldsymbol{\Phi}\,\mathrm{d}x = -\int_{\mathbb{R}^d} v_h\nabla\cdot\boldsymbol{\Phi}\,\mathrm{d}x + \sum_{F\in\mathcal{F}_h}\int_F\{\boldsymbol{\Phi} - \mathcal{I}_h^{\mathrm{b}}\boldsymbol{\Phi}\}\cdot\boldsymbol{n}_F[\![v_h]\!]\,\mathrm{d}s$ for all $\boldsymbol{\Phi} \in \boldsymbol{C}_0^\infty(\mathbb{R}^d)$.)

Solution to exercises

Exercise 38.1 (Elementary dG identities). (i) Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$ be an interface. We have

$$\begin{aligned}\boldsymbol{\sigma}_{|K_l} q_{|K_l} - \boldsymbol{\sigma}_{|K_r} q_{|K_r} &= \frac{1}{2}\boldsymbol{\sigma}_{|K_l} q_{|K_l} - \frac{1}{2}\boldsymbol{\sigma}_{|K_l} q_{|K_r} + \frac{1}{2}\boldsymbol{\sigma}_{|K_l} q_{|K_r} + \frac{1}{2}\boldsymbol{\sigma}_{|K_l} q_{|K_l} \\ &\quad - \frac{1}{2}\boldsymbol{\sigma}_{|K_r} q_{|K_r} + \frac{1}{2}\boldsymbol{\sigma}_{|K_r} q_{|K_l} - \frac{1}{2}\boldsymbol{\sigma}_{|K_r} q_{|K_l} - \frac{1}{2}\boldsymbol{\sigma}_{|K_r} q_{|K_r} \\ &= \frac{1}{2}\boldsymbol{\sigma}_{|K_l}[\![q]\!] + \boldsymbol{\sigma}_{|K_l}\{q\} + \frac{1}{2}\boldsymbol{\sigma}_{|K_r}[\![q]\!] - \boldsymbol{\sigma}_{|K_r}\{q\} \\ &= \{\boldsymbol{\sigma}\}[\![q]\!] + [\![\boldsymbol{\sigma}]\!]\{q\},\end{aligned}$$

and the result follows after observing that $\boldsymbol{n}_{K_l} = -\boldsymbol{n}_{K_r} =: \boldsymbol{n}_F$.
(ii) We proceed as above and obtain that

$$\begin{aligned}\frac{1}{2}(a_l b_l + a_r b_r) &= \frac{1}{2}(\theta_l a_l b_l + \theta_r a_l b_r - \theta_r a_l b_r + \theta_r a_l b_l) + \frac{1}{2}(\theta_r a_r b_r + \theta_l a_r b_l - \theta_l a_r b_l + \theta_l a_r b_r) \\ &= \frac{1}{2}a_l\{b\}_\theta + \frac{1}{2}a_r\{b\}_\theta + \frac{1}{2}(\theta_r a_l - \theta_l a_r)b_l - \frac{1}{2}(\theta_r a_l - \theta_l a_r)b_r \\ &= \{a\}\{b\}_\theta + \frac{1}{4}[\![a]\!]_\theta[\![b]\!].\end{aligned}$$

Exercise 38.2 (Boundary conditions). (i) Integration by parts shows that

$$\sum_{K\in\mathcal{T}_h}\int_D \nabla_h u\cdot\nabla_h w_h \,\mathrm{d}x - \sum_{F\in\mathcal{F}_h^\circ}\int_F \{\nabla_h u\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s - \sum_{F\in\mathcal{F}_h^\partial}\int_F (\nabla_h u\cdot\boldsymbol{n}_K)w_h\,\mathrm{d}s = \int_D f w_h\,\mathrm{d}x.$$

Adding the symmetry term and the penalty term on the interfaces on the left-hand side does not change anything since u is continuous across the interfaces. Notice though that the symmetry and the penalty terms are not zero at the boundary. Hence, we must add them on both sides of the equation, yielding

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}\int_D \nabla_h u\cdot\nabla_h w_h \,\mathrm{d}x - \sum_{F\in\mathcal{F}_h^\circ}\int_F \{\nabla_h u\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s - \sum_{F\in\mathcal{F}_h^\partial}\int_F (\nabla_h u\cdot\boldsymbol{n}_K)w_h\,\mathrm{d}s \\ &- \theta\sum_{F\in\mathcal{F}_h^\circ}\int_F \{\nabla_h w_h\}\cdot\boldsymbol{n}_F[\![u]\!]\,\mathrm{d}s - \theta\sum_{F\in\mathcal{F}_h^\partial}\int_F (\nabla_h w_h\cdot\boldsymbol{n}_K)u\,\mathrm{d}s \\ &+ \sum_{F\in\mathcal{F}_h^\circ}\varpi(h_F)\int_F [\![u]\!][\![w_h]\!]\,\mathrm{d}s + \sum_{F\in\mathcal{F}_h^\partial}\varpi(h_F)\int_F u w_h\,\mathrm{d}s \\ &= \int_D f w_h\,\mathrm{d}x - \theta\sum_{F\in\mathcal{F}_h^\partial}\int_F (\nabla_h w_h\cdot\boldsymbol{n}_K)u\,\mathrm{d}s + \sum_{F\in\mathcal{F}_h^\partial}\varpi(h_F)\int_F u w_h\,\mathrm{d}s,\end{aligned}$$

where the value of $\theta \in \{-1, 0, 1\}$ depends on the method that is chosen (NIP, IIP, SIP). Now, we replace $u_{|\partial D}$ by g on the right-hand side and we regroup the boundary and interface integrals

using the usual convention about jumps and averages at the boundary. This leads to

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}\int_D \nabla_h u\cdot\nabla_h w_h\,\mathrm{d}x - \sum_{F\in\mathcal{F}_h}\int_F \{\nabla_h u\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s\\ &-\theta\sum_{F\in\mathcal{F}_h}\int_F \{\nabla_h w_h\}\cdot\boldsymbol{n}_F[\![u]\!]\,\mathrm{d}s + \sum_{F\in\mathcal{F}_h}\varpi(h_F)\int_F [\![u]\!][\![w_h]\!]\,\mathrm{d}s\\ &=\int_D f w_h\,\mathrm{d}x - \theta\sum_{F\in\mathcal{F}_h^\partial}\int_F(\nabla_h w_h\cdot\boldsymbol{n}_K)g\,\mathrm{d}s + \sum_{F\in\mathcal{F}_h^\partial}\varpi(h_F)\int_F g w_h\,\mathrm{d}s.\end{aligned}$$

Thus, the exact consistency property $a_h^\theta(u,w_h)=\ell_h^{\theta,\mathrm{nD}}(w_h)$ holds true for all $w_h\in V_h$ if we set

$$\ell_h^{\theta,\mathrm{nD}}(w_h) := \int_D f w_h\,\mathrm{d}x - \theta\sum_{F\in\mathcal{F}_h^\partial}\int_F(\nabla_h w_h\cdot\boldsymbol{n}_K)g\,\mathrm{d}s + \sum_{F\in\mathcal{F}_h^\partial}\varpi(h_F)\int_F g w_h\,\mathrm{d}s.$$

(ii) We proceed as above for the Robin boundary condition. The only difference is that we do not add the symmetry term and the penalty term at the boundary. This leads to

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}\int_D \nabla_h u\cdot\nabla_h w_h\,\mathrm{d}x - \sum_{F\in\mathcal{F}_h^\circ}\int_F \{\nabla_h u\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s - \sum_{F\in\mathcal{F}_h^\partial}\int_F(\nabla_h u\cdot\boldsymbol{n}_K)w_h\,\mathrm{d}s\\ &-\theta\sum_{F\in\mathcal{F}_h^\circ}\int_F \{\nabla_h w_h\}\cdot\boldsymbol{n}_F[\![u]\!]\,\mathrm{d}s + \sum_{F\in\mathcal{F}_h^\circ}\varpi(h_F)\int_F [\![u]\!][\![w_h]\!]\,\mathrm{d}s = \int_D f w_h\,\mathrm{d}x.\end{aligned}$$

We conclude by replacing $\nabla_h u\cdot\boldsymbol{n}_K$ at the boundary by $-\gamma u+g$, leading to

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}\int_D \nabla_h u\cdot\nabla_h w_h\,\mathrm{d}x - \sum_{F\in\mathcal{F}_h^\circ}\int_F \{\nabla_h u\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s + \sum_{F\in\mathcal{F}_h^\partial}\int_F \gamma u w_h\,\mathrm{d}s\\ &-\theta\sum_{F\in\mathcal{F}_h^\circ}\int_F \{\nabla_h w_h\}\cdot\boldsymbol{n}_F[\![u]\!]\,\mathrm{d}s + \sum_{F\in\mathcal{F}_h^\circ}\varpi(h_F)\int_F [\![u]\!][\![w_h]\!]\,\mathrm{d}s = \int_D f w_h\,\mathrm{d}x + \sum_{F\in\mathcal{F}_h^\partial}\int_F g w_h\,\mathrm{d}s.\end{aligned}$$

Therefore, once again the exact consistency property, i.e.,

$$a_h^{\theta,\mathrm{Rb}}(u,w_h)=\ell_h(w_h)+\sum_{F\in\mathcal{F}_h^\partial}\int_F g w_h\,\mathrm{d}s =: \ell_h^{\mathrm{Rb}}(w_h),\qquad \forall w_h\in V_h,$$

holds true if we set

$$\begin{aligned}a_h^{\theta,\mathrm{Rb}}(v,w_h) := &\sum_{K\in\mathcal{T}_h}\int_D \nabla_h v\cdot\nabla_h w_h\,\mathrm{d}x\\ &-\sum_{F\in\mathcal{F}_h^\circ}\int_F \{\nabla_h v\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s - \theta\sum_{F\in\mathcal{F}_h^\circ}\int_F \{\nabla_h w_h\}\cdot\boldsymbol{n}_F[\![v]\!]\,\mathrm{d}s\\ &+\sum_{F\in\mathcal{F}_h^\partial}\int_F \gamma v w_h\,\mathrm{d}s + \sum_{F\in\mathcal{F}_h^\circ}\varpi(h_F)\int_F [\![v]\!][\![w_h]\!]\,\mathrm{d}s.\end{aligned}$$

Exercise 38.3 (L^2-estimate). Let $e:=u-u_h$. Let us set $V_\sharp := V_{\mathrm{s}}+V_h$, $Z_{\mathrm{s}} := H^{1+s}(D)\cap H_0^1(D)$, $Y_h := V_h\cap H_0^1(D)$, $Z_\sharp := Z_{\mathrm{s}}+Y_h$ equipped with the same norm as $V_\sharp$. Consider the bilinear form

$$a_\sharp(v,w) := (\nabla_h v,\nabla_h w)_{\boldsymbol{L}^2(D)} - \sum_{F\in\mathcal{F}_h}\int_F [\![v]\!]\{\nabla_h w\}\cdot\boldsymbol{n}_F\,\mathrm{d}s.$$

Notice that $a_\sharp$ is bounded on $V_\sharp \times Z_\sharp$, and that for all $y_h \in Y_h$,

$$a_\sharp(u, y_h) = (f, y_h)_{L^2(D)} = \ell_h(y_h) = a_h(u_h, y_h) = a_\sharp(u_h, y_h),$$

since $Y_h \subset H^1_0(D)$ and $[\![u]\!] = 0$ for all $F \in \mathcal{F}_h$. Hence, the Galerkin orthogonality property holds true for $a_\sharp$. We can therefore apply the abstract nonconforming estimate from Lemma 36.14 which yields

$$\|e\|_{L^2(D)} \le \left(\frac{\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'}}{\|e\|_{L^2(D)}} + \inf_{y_h \in Y_h} \frac{\|\nabla(\zeta_e - y_h)\|_{\boldsymbol{L}^2(D)}}{\|e\|_{L^2(D)}} \right) \|e\|_{V_\sharp},$$

where $\zeta_e \in H^1_0(D)$ is the adjoint solution associated with the error e (recall that $\|\zeta\|_{H^{1+s}(D)} \le c\ell_D^2 \|e\|_{L^2(D)}$) and where the two terms between parentheses are the adjoint consistency error and the interpolation error on the adjoint solution. Let us first bound the adjoint consistency error. The definition of $\delta^{\mathrm{adj}}(\zeta_e)$ implies that for all $v \in V_\sharp$,

$$\langle \delta^{\mathrm{adj}}(\zeta_e), v \rangle_{V_\sharp', V_\sharp} := -(v, \Delta\zeta_e)_{L^2(D)} - a_\sharp(v, \zeta_e) = 0,$$

where we used that $[\![\nabla\zeta_e]\!]_F = \mathbf{0}$ for all $F \in \mathcal{F}_h^\circ$ since $s > \frac{1}{2}$, i.e., exact adjoint consistency holds true. To bound the interpolation error on the adjoint solution, we consider the quasi-interpolation operator $\mathcal{I}_{h0}^{\mathrm{g,av}}$ from §22.4 and deduce that

$$\begin{aligned}
\inf_{y_h \in Y_h} \|\zeta_e - y_h\|_{Z_\sharp} &\le \|\zeta_e - \mathcal{I}_{h0}^{\mathrm{g,av}}(\zeta_e)\|_{Z_\sharp} \\
&\le c\, h^s |\zeta_e|_{H^{1+s}(D)} \\
&\le c\, h^s \ell_D^{-1-s} \|\zeta_e\|_{H^{1+s}(D)} \\
&\le c\, c_{\mathrm{smo}}\, h^s \ell_D^{1-s} \|e\|_{L^2(D)},
\end{aligned}$$

where we used the approximation properties of $\mathcal{I}_{h0}^{\mathrm{g,av}}$ from Theorem 22.14. *Note*: it is also possible to estimate the interpolation error using any of the operators from §18.3 (e.g., the L^2-orthogonal projection) by considering the bilinear form

$$\begin{aligned}
a_\sharp(v, w) = (\nabla_h v, \nabla_h w)_{\boldsymbol{L}^2(D)} &- \sum_{F \in \mathcal{F}_h} \int_F \{\nabla_h v\} \cdot \boldsymbol{n}_F [\![w]\!] \,\mathrm{d}s \\
&- \sum_{F \in \mathcal{F}_h} \int_F [\![v]\!] \{\nabla_h w\} \cdot \boldsymbol{n}_F \,\mathrm{d}s + \sum_{F \in \mathcal{F}_h} \frac{\varpi_0}{h_F} \int_F [\![v]\!][\![w]\!] \,\mathrm{d}s.
\end{aligned}$$

Exercise 38.4 (Local lifting). We observe that

$$\|\boldsymbol{\mathcal{L}}_F^l(\varphi)\|_{L^2(D_F)}^2 = \int_F \{\boldsymbol{\mathcal{L}}_F^l(\varphi)\} \cdot \boldsymbol{n}_F \varphi \,\mathrm{d}s.$$

Using the Cauchy–Schwarz inequality together with the fact that $\{w\} = |\mathcal{T}_F|^{-1} \sum_{K \in \mathcal{T}_F} w_{|K}$ for every function w, we infer that

$$\begin{aligned}
\|\boldsymbol{\mathcal{L}}_F^l(\varphi)\|_{L^2(D_F)}^2 &\le \|\{\boldsymbol{\mathcal{L}}_F^l(\varphi)\}\|_{\boldsymbol{L}^2(F)} \|[\![\varphi]\!]\|_{L^2(F)} \\
&= h_F^{\frac{1}{2}} \|\{\boldsymbol{\mathcal{L}}_F^l(\varphi)\}\|_{\boldsymbol{L}^2(F)} \times h_F^{-\frac{1}{2}} \|[\![\varphi]\!]\|_{L^2(F)} \\
&\le \frac{1}{|\mathcal{T}_F|} \sum_{K \in \mathcal{T}_F} h_F^{\frac{1}{2}} \|\boldsymbol{\mathcal{L}}_F^l(\varphi)_{|K}\|_{\boldsymbol{L}^2(F)} \times h_F^{-\frac{1}{2}} \|[\![\varphi]\!]\|_{L^2(F)} \\
&\le \frac{1}{|\mathcal{T}_F|} \sum_{K \in \mathcal{T}_F} c_{\mathrm{dt}} \|\boldsymbol{\mathcal{L}}_F^l(\varphi)\|_{\boldsymbol{L}^2(K)} \times h_F^{-\frac{1}{2}} \|[\![\varphi]\!]\|_{L^2(F)} \\
&\le c_{\mathrm{dt}} \|\boldsymbol{\mathcal{L}}_F^l(\varphi)\|_{L^2(D_F)} \times h_F^{-\frac{1}{2}} \|[\![\varphi]\!]\|_{L^2(F)},
\end{aligned}$$

since $\sum_{K\in\mathcal{T}_F}\|\boldsymbol{\mathcal{L}}_F^l(\varphi)\|_{\boldsymbol{L}^2(K)}\le|\mathcal{T}_F|^{\frac12}\|\boldsymbol{\mathcal{L}}_F^l(\varphi)\|_{L^2(D_F)}$ and $|\mathcal{T}_F|\ge1$.

Exercise 38.5 (Local formulation). The local formulations take the form

$$\int_K\boldsymbol{G}_h(u_h)\cdot\nabla q\,\mathrm{d}x+\sum_{F\in\mathcal{F}_K}(\boldsymbol{n}_K\cdot\boldsymbol{n}_F)\int_F\widehat{\Phi}_F(u_h)q\,\mathrm{d}s=\int_K fq\,\mathrm{d}x,$$

with

$$\begin{aligned}
\boldsymbol{G}_h^{\mathrm{OBB}}(u_h)&:=\nabla_h u_h+\boldsymbol{\mathcal{L}}_h^l([\![u_h]\!]),&\qquad\widehat{\Phi}_F^{\mathrm{OBB}}(u_h)&:=-\{\nabla_h u_h\}\cdot\boldsymbol{n}_F,\\
\boldsymbol{G}_h^{\mathrm{NIP}}(u_h)&:=\nabla_h u_h+\boldsymbol{\mathcal{L}}_h^l([\![u_h]\!]),&\qquad\widehat{\Phi}_F^{\mathrm{NIP}}(u_h)&:=-\{\nabla_h u_h\}\cdot\boldsymbol{n}_F+\varpi(h_F)[\![u_h]\!],\\
\boldsymbol{G}_h^{\mathrm{IIP}}(u_h)&:=\nabla_h u_h,&\qquad\widehat{\Phi}_F^{\mathrm{IIP}}(u_h)&:=-\{\nabla_h u_h\}\cdot\boldsymbol{n}_F+\varpi(h_F)[\![u_h]\!].
\end{aligned}$$

Exercise 38.6 (Extending (38.25)**).** We observe that

$$\begin{aligned}
\tilde{a}_h(v,w_h)-a_h(v,w_h)&=-\int_{\mathbb{R}^d}\nabla_h v\cdot\boldsymbol{\mathcal{L}}_h^l([\![w_h]\!])\,\mathrm{d}x+\sum_{F\in\mathcal{F}_h}\int_F\{\nabla_h v\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s\\
&=-\int_{\mathbb{R}^d}\mathcal{I}_h^{\mathrm{b}}(\nabla_h v)\cdot\boldsymbol{\mathcal{L}}_h^l([\![w_h]\!])\,\mathrm{d}x+\sum_{F\in\mathcal{F}_h}\int_F\{\nabla_h v\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s\\
&=\sum_{F\in\mathcal{F}_h}\int_F\{\nabla_h v-\mathcal{I}_h^{\mathrm{b}}(\nabla_h v)\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s,
\end{aligned}$$

where we used that $\boldsymbol{\mathcal{L}}_h^l([\![v_h]\!])\in P_l^{\mathrm{b}}(\mathcal{T}_h)$ and the definition of $\boldsymbol{\mathcal{L}}_h^l([\![v_h]\!])$.

Exercise 38.7 (Discrete gradient). (i) The sequence $(\mathfrak{G}_h^l(v_h))_{h\in\mathcal{H}}$ is uniformly bounded in $\boldsymbol{L}^2(D)$ since the triangle inequality and (38.22b) imply that

$$\|\mathfrak{G}_h^l(v_h)\|_{\boldsymbol{L}^2(D)}\le\|\nabla_h v_h\|_{\boldsymbol{L}^2(D)}+\|\boldsymbol{\mathcal{L}}_h^l([\![v_h]\!])\|_{\boldsymbol{L}^2(D)}\le\max(1,n_\partial^{\frac12}c_{\mathrm{dt}})\sqrt{2}\|v_h\|_{V_h}.$$

Since $\boldsymbol{L}^2(D)$ is a Hilbert space (which is a reflexive Banach space), Theorem C.23 implies that there is $\boldsymbol{G}\in\boldsymbol{L}^2(D)$ s.t. $\mathfrak{G}_h^l(v_h)$ weakly converges to $\boldsymbol{G}$ in $\boldsymbol{L}^2(D)$ as $h\to0$.
(ii) Let now $\boldsymbol{\Phi}\in\boldsymbol{C}_0^\infty(\mathbb{R}^d)$. We have

$$\begin{aligned}
\int_{\mathbb{R}^d}\mathfrak{G}_h^l(v_h)\cdot\boldsymbol{\Phi}\,\mathrm{d}x&=\int_{\mathbb{R}^d}\nabla_h v_h\cdot\boldsymbol{\Phi}\,\mathrm{d}x-\int_{\mathbb{R}^d}\boldsymbol{\mathcal{L}}_h^l([\![v_h]\!])\cdot\boldsymbol{\Phi}\,\mathrm{d}x\\
&=\int_{\mathbb{R}^d}\nabla_h v_h\cdot\boldsymbol{\Phi}\,\mathrm{d}x-\int_{\mathbb{R}^d}\boldsymbol{\mathcal{L}}_h^l([\![v_h]\!])\cdot\mathcal{I}_h^{\mathrm{b}}(\boldsymbol{\Phi})\,\mathrm{d}x\\
&=-\int_{\mathbb{R}^d}v_h\nabla\cdot\boldsymbol{\Phi}\,\mathrm{d}x+\sum_{F\in\mathcal{F}_h}\int_F[\![v_h]\!]\boldsymbol{\Phi}\cdot\boldsymbol{n}_F\,\mathrm{d}s-\int_{\mathbb{R}^d}\boldsymbol{\mathcal{L}}_h^l([\![v_h]\!])\cdot\mathcal{I}_h^{\mathrm{b}}(\boldsymbol{\Phi})\,\mathrm{d}x\\
&=-\int_{\mathbb{R}^d}v_h\nabla\cdot\boldsymbol{\Phi}\,\mathrm{d}x+\sum_{F\in\mathcal{F}_h}\int_F[\![v_h]\!]\{\boldsymbol{\Phi}-\mathcal{I}_h^{\mathrm{b}}(\boldsymbol{\Phi})\}\cdot\boldsymbol{n}_F\,\mathrm{d}s,
\end{aligned}$$

where we used elementwise integration by parts and proceeded as in Exercise 38.6. Let $\mathfrak{T}_1,\mathfrak{T}_2$ denote the two terms on the right-hand side. The convergence of v_h to v in $L^2(\mathbb{R}^d)$ implies that $\mathfrak{T}_1\to-\int_{\mathbb{R}^d}v\nabla\cdot\boldsymbol{\Phi}\,\mathrm{d}x$. Moreover, the Cauchy–Schwarz inequality and the regularity of the mesh sequence lead to

$$|\mathfrak{T}_2|\le|v_h|_{\mathrm{J}}\left(\sum_{K\in\mathcal{T}_h}h_K\|\boldsymbol{\Phi}-\mathcal{I}_h^{\mathrm{b}}(\boldsymbol{\Phi})\|_{L^2(\partial K)}^2\right)^{\frac12}.$$

Using the Poincaré–Steklov inequality, we infer that

$$|\mathfrak{T}_2| \leq |v_h|_{\mathrm{J}} \left(\sum_{K \in \mathcal{T}_h} h_K^2 |\boldsymbol{\Phi}|^2_{\boldsymbol{H}^1(K)} \right)^{\frac{1}{2}} \leq h |v_h|_{\mathrm{J}} |\boldsymbol{\Phi}|_{\boldsymbol{H}^1(D)} \to 0,$$

since $|v_h|_{\mathrm{J}}$ is uniformly bounded w.r.t. $h \in \mathcal{H}$. Letting $h \to 0$ in the above equality and using the weak convergence of $\mathfrak{G}_h^l(v_h)$ to $\boldsymbol{G}$, we infer that

$$\int_{\mathbb{R}^d} \boldsymbol{G} \cdot \boldsymbol{\Phi} \, \mathrm{d}x = - \int_{\mathbb{R}^d} v \nabla \cdot \boldsymbol{\Phi} \, \mathrm{d}x.$$

This shows that $v \in H^1(\mathbb{R}^d)$ with $\nabla v = \boldsymbol{G}$. Since v has been extended by zero outside D, we infer that $\gamma^{\mathrm{g}}(v) = 0$, i.e., $v \in H_0^1(D)$.

Chapter 39

Hybrid high-order method

Exercises

Exercise 39.1 (Stabilization). Prove that $\hat{a}_K(\hat{v}_K,\hat{v}_K)$ is equivalent to $\|\nabla r_K\|^2_{\boldsymbol{L}^2(K)}+\hat{\theta}_K(\hat{v}_K,\hat{v}_K)$ for all $\hat{v}_K\in\hat{V}_K^k$, with $r_K:=\mathsf{R}(\hat{v}_K)$ and

$$\hat{\theta}_K(\hat{v}_K,\hat{v}_K):=h_K^{-2}\|v_K-\Pi_K^k(r_K)\|^2_{L^2(K)}+h_K^{-1}\|v_{\partial K}-\Pi_{\partial K}^k(r_K)\|^2_{L^2(\partial K)}.$$

(*Hint*: note that $\mathsf{S}(\hat{v}_K)=\Pi_{\partial K}^k(v_K-\Pi_K^k(r_K))_{|\partial K}-(v_{\partial K}-\Pi_{\partial K}^k(r_K))$, and to bound $\hat{a}_K(\hat{v}_K,\hat{v}_K)$ from below, prove that $\hat{\theta}_K(\hat{v}_K,\hat{v}_K)^{\frac12}\le c\,h_K^{-1}\|v_K-r_K\|_{L^2(K)}+h_K^{-\frac12}\|\mathsf{S}(\hat{v}_K)\|_{L^2(\partial K)}$, then invoke the Poincaré–Steklov inequality, the triangle inequality, and the lower bound from Lemma 39.2.)

Exercise 39.2 (Finite element viewpoint). Let $\mathcal{V}_K^k$ be defined in (39.10). Let $\mathcal{E}_K:H^1(K)\to V_K^{k+1}$ be the elliptic projection and set $\delta:=v-\mathcal{E}_K(v)$ for all $v\in\mathcal{V}_K^k$. (i) Prove that

$$h_K^{-1}\|\Pi_K^k(\delta)\|_{L^2(K)}\le c\left(\|\nabla\mathcal{E}_K(v)\|_{\boldsymbol{L}^2(K)}+h_K^{-\frac12}\|\mathsf{S}(\hat{\mathcal{I}}_K^k(v))\|_{L^2(\partial K)}\right).$$

(*Hint*: use the Poincaré–Steklov inequality in K and the lower bound from Lemma 39.2.) (ii) Prove that

$$\|\nabla\delta\|_{\boldsymbol{L}^2(K)}\le c\left(\|\nabla\mathcal{E}_K(v)\|_{\boldsymbol{L}^2(K)}+h_K^{-\frac12}\|\mathsf{S}(\hat{\mathcal{I}}_K^k(v))\|_{L^2(\partial K)}\right).$$

(*Hint*: integrate by parts $\|\nabla\delta\|^2_{\boldsymbol{L}^2(K)}$ and accept as a fact that a discrete trace inequality and an inverse inequality are valid on $\mathcal{V}_K^k$, then use that $\mathsf{S}(\hat{\mathcal{I}}_K^k(v))=\Pi_{\partial K}^k(\Pi_K^k(\delta)_{|\partial K})-\Pi_{\partial K}^k(\delta_{|\partial K})$.) (iii) Let $\mathfrak{a}_K(v,w):=(\nabla\mathcal{E}_K(v),\nabla\mathcal{E}_K(w))_{\boldsymbol{L}^2(K)}+h_K^{-1}(\mathsf{S}(\hat{\mathcal{I}}_K^k(v)),\mathsf{S}(\hat{\mathcal{I}}_K^k(v)))_{L^2(\partial K)}$ on $\mathcal{V}_K^k\times\mathcal{V}_K^k$. Prove that $\mathfrak{a}_K(v,v)\ge c\|\nabla v\|^2_{\boldsymbol{L}^2(K)}$ with $c>0$.

Exercise 39.3 (Elliptic projection). Prove the second bound in Theorem 39.17. (*Hint*: introduce the L^2-orthogonal projection Π_K^{k+1}.)

Exercise 39.4 (Reconstruction). (i) Let $\mathsf{G}:\hat{V}_K^k\to\boldsymbol{V}_K^k:=\mathbb{P}_{k,d}\circ\boldsymbol{T}_K^{-1}$ be s.t. $(\mathsf{G}(\hat{v}_K),\boldsymbol{q})_{\boldsymbol{L}^2(K)}=-(v_K,\nabla\cdot\boldsymbol{q})_{L^2(K)}+(v_{\partial K},\boldsymbol{n}_K\cdot\boldsymbol{q})_{L^2(\partial K)}$ for all $\boldsymbol{q}\in\boldsymbol{V}_K^k$. Prove that $\Pi_{\nabla V_K^{k+1}}\mathsf{G}=\nabla\mathsf{R}$, where $\Pi_{\nabla V_K^{k+1}}$ is the $\boldsymbol{L}^2$-orthogonal projection onto ∇V_K^{k+1}. (ii) Let $\mathsf{G}_{\mathrm{RT}}:\hat{V}_K^k\to\boldsymbol{V}_K^k:=(\boldsymbol{\psi}_K^{\mathrm{d}})^{-1}(\mathbb{RT}_{k,d})$ be s.t. $(\mathsf{G}_{\mathrm{RT}}(\hat{v}_K),\boldsymbol{q})_{\boldsymbol{L}^2(K)}=-(v_K,\nabla\cdot\boldsymbol{q})_{L^2(K)}+(v_{\partial K},\boldsymbol{n}_K\cdot\boldsymbol{q})_{L^2(\partial K)}$ for all $\boldsymbol{q}\in\boldsymbol{V}_K^k$, where $\boldsymbol{\psi}_K^{\mathrm{d}}$ is the contravariant Piola transformation defined in (9.9c), and $\mathbb{RT}_{k,d}$ is the Raviart–Thomas polynomial

space. Prove that $\|\mathsf{G}_{\mathrm{RT}}(\hat{v}_K)\|_{\boldsymbol{L}^2(K)} \geq c|\hat{v}_K|_{\hat{V}_K^k}$ with $c > 0$. (*Hint*: use the dofs of the Raviart–Thomas element; see John et al. [30] for the seminal idea in the context of dG methods.)

Exercise 39.5 ($k = 0$). (i) Derive the HHO method in 1D for $k = 0$, as well as the global transmission problem. (ii) Prove that, in dimension $d \geq 2$ for $k = 0$, $\mathsf{R}(\hat{v}_K)(\boldsymbol{x}) = v_K + \sum_{F\in\mathcal{F}_K} \frac{|F|}{|K|}(v_F - v_K)\boldsymbol{n}_{K|F}\cdot(\boldsymbol{x} - \boldsymbol{x}_K)$ for all $\boldsymbol{x} \in K$, with $v_F := v_{\partial K|F}$ for all $F \in \mathcal{F}_K$, and $\boldsymbol{x}_K$ is the barycenter of K, and $\mathsf{S}(\hat{v}_K)_{|F} = v_K - v_F - \nabla\mathsf{R}(\hat{v}_K)\cdot(\boldsymbol{x}_K - \boldsymbol{x}_F)$, where $\boldsymbol{x}_F$ is the barycenter of F for all $F \in \mathcal{F}_K$ (*Hint*: any function $q \in \mathbb{P}_{1,d} \circ \boldsymbol{T}_K^{-1}$ is of the form $q(\boldsymbol{x}) = q_K + \boldsymbol{G}_q\cdot(\boldsymbol{x} - \boldsymbol{x}_K)$, where $q_K := q(\boldsymbol{x}_K)$ is the mean value of q over K and $\boldsymbol{G}_q := \nabla q$, and use also (7.1).)

Exercise 39.6 (Transmission problem). (i) Prove the converse statement in Proposition 39.10. (*Hint*: write $\hat{w}_K = (w_K - U_{w_{\partial K}}, 0) + (U_{w_{\partial K}}, w_{\partial K})$.) (ii) Justify Remark 39.11. (*Hint*: for the converse statement show that $a_K(u, w) - \ell_K(w) = a_K(U_{\lambda_{\partial K}}, U_\mu) - \ell_K(U_\mu)$ with $\mu := w_{\partial K}$.) (iii) Adapt the statement if a_K is nonsymmetric. (*Hint*: consider $U_\lambda^* \in H^1(K)$ s.t. $U^*_{\lambda|\partial K} = \lambda$ and $a_K(\psi, U_\lambda^*) = 0$ for all $\psi \in H_0^1(K)$.) (iv) Prove (39.23).

Exercise 39.7 (HDG). Consider the HDG method. Assume the following: if $(v_K, \mu_{\partial K}) \in V_K \times V_{\partial K}$ with $V_{\partial K} := \prod_{F\in\mathcal{F}_K} V_F$ is s.t. $(\tau_{\partial K}(v_{K|\partial K} - \mu_{\partial K}), v_{K|\partial K} - \mu_{\partial K})_{L^2(\partial K)} = 0$ and $(v_K, \nabla\cdot\boldsymbol{\tau}_K)_{L^2(K)} - (\mu_{\partial K}, \boldsymbol{\tau}_K\cdot\boldsymbol{n}_K)_{L^2(\partial K)} = 0$ for all $\boldsymbol{\tau}_K \in \boldsymbol{S}_K$, then v_K and $\mu_{\partial K}$ are constant functions taking the same value. Prove that the discrete problem (39.25) is well-posed. (*Hint*: derive an energy identity.)

Exercise 39.8 (Space Λ). Let Λ be defined in (39.21). Recall that the trace map $\gamma^{\mathrm{g}}_{\partial K} : H^1(K) \to H^{\frac12}(\partial K)$ is surjective. (i) Prove that there are constants $0 < c_1 \leq c_2$ s.t. $c_1\|\nabla U_\mu\|_{\boldsymbol{L}^2(K)} \leq |\mu|_{H^{\frac12}(\partial K)} \leq c_2\|\nabla U_\mu\|_{\boldsymbol{L}^2(K)}$ for all $\mu \in H^{\frac12}(\partial K)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$. (*Hint*: prove first the bounds on the reference cell $\widehat{K}$.) (ii) Set $\|\lambda\|_\Lambda^2 := \sum_{K\in\mathcal{T}_h} |\lambda_{\partial K}|^2_{H^{\frac12}(\partial K)}$. Verify that $\|\cdot\|_\Lambda$ indeed defines a norm on Λ, and that Λ is a Hilbert space. (*Hint*: for all $\lambda \in \Lambda$, consider the function $U_\lambda : D \to \mathbb{R}$ s.t. $U_{\lambda|K} := U_{\lambda_{\partial K}}$ for all $K \in \mathcal{T}_h$, and prove that $U_\lambda \in H_0^1(D)$.)

Exercise 39.9 (Liftings, 1D). Consider a uniform mesh of $D := (0, 1)$ with nodes $x_i := ih$, $i := \frac{1}{I+1}$ for all $i \in \{0{:}(I{+}1)\}$. Consider the PDE $-u'' = f$ in D with $u(0) = u(1) = 0$. (i) Prove that (39.22) amounts to $\mathcal{A}X = B$ with $\mathcal{A} = h^{-1}\,\mathrm{tridiag}(-1, 2, -1)$, $X_i = \lambda_i$, and $B_i = \int_{x_{i-1}}^{x_{i+1}} \varphi_i f \,\mathrm{d}s$ for all $i \in \{1{:}I\}$. (*Hint*: prove that U_λ is affine on every cell $K_i = [x_{i-1}, x_i]$.) Prove that $\lambda_i = u(x_i)$. (*Hint*: write $f = -u''$ and integrate by parts. This remarkable fact only happens in 1D.) (ii) Let $k \geq 2$. For all $m \geq 1$, set $\phi_m := (2(2m + 1))^{-\frac12}(L_{m+1} - L_{m-1})$, where L_m is the Legendre polynomial of degree m (see §6.1). Verify that $\{\phi_m\}_{m\in\{1:k-1\}}$ is a basis of $\mathbb{P}_k^\circ := \{p \in \mathbb{P}_k \mid p(\pm1) = 0\}$. Prove that $U_{f_{|\widehat{K}}}(x) = \int_{\widehat{K}} G(x, s)f(s) \,\mathrm{d}s$ on $\widehat{K} := [-1, 1]$ with the discrete Green's function $G(x, s) := \sum_{m\in\{1:k-1\}} \phi_m(x)\phi_m(s)$. (*Hint*: observe that $\phi_m' = L_m$.) Infer the expression of $U_{f_{|K_i}}$ for every cell K_i.

Solution to exercises

Exercise 39.1 (Stabilization). Using the hint and the triangle inequality leads to

$$\begin{aligned}\|\mathsf{S}(\hat{v}_K)\|_{L^2(\partial K)} &\leq \|\Pi^k_{\partial K}(v_K - \Pi^k_K(r_K))_{|\partial K}\|_{L^2(\partial K)} + \|v_{\partial K} - \Pi^k_{\partial K}(r_K)\|_{L^2(\partial K)}\\ &\leq c\,h_K^{-\frac12}\|v_K - \Pi^k_K(r_K)\|_{L^2(K)} + \|v_{\partial K} - \Pi^k_{\partial K}(r_K)\|_{L^2(\partial K)}\\ &\leq c' h_K^{\frac12}\hat{\theta}_K(\hat{v}_K, \hat{v}_K)^{\frac12},\end{aligned}$$

where we used the L^2-stability of $\Pi_{\partial K}^k$ and a discrete trace inequality. Hence, we have

$$\begin{aligned}\hat a_K(\hat v_K,\hat v_K) &= \|\nabla r_K\|_{\boldsymbol{L}^2(K)}^2 + h_K^{-1}\|\mathsf{S}(\hat v_K)\|_{L^2(\partial K)}^2\\ &\le \|\nabla r_K\|_{\boldsymbol{L}^2(K)}^2 + c\,\hat\theta_K(\hat v_K,\hat v_K).\end{aligned}$$

Let us now prove the converse bound. Using again the identity from the hint and the triangle inequality, we infer that

$$\begin{aligned}\hat\theta_K(\hat v_K,\hat v_K)^{\frac12} \le{}& h_K^{-1}\|v_K-\Pi_K^k(r_K)\|_{L^2(K)} + h_K^{-\frac12}\|\mathsf{S}(\hat v_K)\|_{L^2(\partial K)}\\ &+ h_K^{-\frac12}\|\Pi_{\partial K}^k(v_K-\Pi_K^k(r_K))_{|\partial K}\|_{L^2(\partial K)}.\end{aligned}$$

Rearranging the terms, using the L^2-stability of $\Pi_{\partial K}^k$ and a discrete trace inequality leads to

$$\hat\theta_K(\hat v_K,\hat v_K)^{\frac12} \le c\,h_K^{-1}\|v_K-\Pi_K^k(r_K)\|_{L^2(K)} + h_K^{-\frac12}\|\mathsf{S}(\hat v_K)\|_{L^2(\partial K)}.$$

Since $v_K-\Pi_K^k(r_K)=\Pi_K^k(v_K-r_K)$, we invoke the L^2-stability of Π_K^k to obtain

$$\hat\theta_K(\hat v_K,\hat v_K)^{\frac12} \le c\,h_K^{-1}\|v_K-r_K\|_{L^2(K)} + h_K^{-\frac12}\|\mathsf{S}(\hat v_K)\|_{L^2(\partial K)}.$$

Owing to the Poincaré–Steklov inequality (recall that v_K-r_K has zero mean value on K), we infer that

$$\hat\theta_K(\hat v_K,\hat v_K)^{\frac12} \le c\,\|\nabla(v_K-r_K)\|_{\boldsymbol{L}^2(K)} + h_K^{-\frac12}\|\mathsf{S}(\hat v_K)\|_{L^2(\partial K)}.$$

The triangle inequality and the lower bound from Lemma 39.2 imply that

$$\hat\theta_K(\hat v_K,\hat v_K)^{\frac12} \le c\,\Big(\|\nabla r_K\|_{\boldsymbol{L}^2(K)} + h_K^{-\frac12}\|\mathsf{S}(\hat v_K)\|_{L^2(\partial K)}\Big).$$

This proves that $\hat\theta_K(\hat v_K,\hat v_K)^{\frac12} \le c\hat a_K(\hat v_K,\hat v_K)^{\frac12}$, and since $\|\nabla r_K\|_{\boldsymbol{L}^2(K)} \le \hat a_K(\hat v_K,\hat v_K)^{\frac12}$, we conclude that $\|\nabla r_K\|_{\boldsymbol{L}^2(K)}^2 + \hat\theta_K(\hat v_K,\hat v_K) \le c\hat a_K(\hat v_K,\hat v_K)$.

Exercise 39.2 (Finite element viewpoint). (i) Let $v\in\mathcal{V}_K^k$ and let us set $\delta := v-\mathcal{E}_K(v)$. The L^2-stability of Π_K^k and the Poincaré–Steklov inequality (recall that δ has zero mean value on K) imply that

$$\|\Pi_K^k(\delta)\|_{L^2(K)} \le \|\delta\|_{L^2(K)} \le c\,h_K\|\nabla\delta\|_{\boldsymbol{L}^2(K)}.$$

The triangle inequality and the lower bound from Lemma 39.2 yield

$$h_K^{-1}\|\Pi_K^k(\delta)\|_{L^2(K)} \le c\,\big(\|\nabla\mathcal{E}_K(v)\|_{\boldsymbol{L}^2(K)} + h_K^{-\frac12}\|\mathsf{S}(\hat{\mathcal{I}}_K^k(v))\|_{L^2(\partial K)}\big).$$

(ii) Integrating by parts and using the definition of $\mathcal{V}_K^k$, we infer that

$$\|\nabla\delta\|_{\boldsymbol{L}^2(K)}^2 = -(\Pi_K^k(\delta),\Delta\delta)_{L^2(K)} + (\Pi_{\partial K}^k(\delta_{|\partial K}),\boldsymbol{n}_K\cdot\nabla\delta)_{L^2(\partial K)}.$$

Invoking the Cauchy–Schwarz inequality, together with a discrete trace inequality and an inverse inequality in $\mathcal{V}_K^k$, leads to

$$\|\nabla\delta\|_{\boldsymbol{L}^2(K)} \le c\,\big(h_K^{-1}\|\Pi_K^k(\delta)\|_{L^2(K)} + h_K^{-\frac12}\|\Pi_{\partial K}^k(\delta_{|\partial K})\|_{L^2(\partial K)}\big),$$

where c is uniform w.r.t. $K\in\mathcal{T}_h$, $h\in\mathcal{H}$, and $v\in\mathcal{V}_K^k$. Recalling the definition of the stabilization operator, rearranging the terms, and recalling that $\mathcal{E}_K=\mathsf{R}\circ\hat{\mathcal{I}}_K^k$, we infer that

$$\mathsf{S}(\hat{\mathcal{I}}_K^k(v)) = \Pi_{\partial K}^k(\Pi_K^k(\delta)_{|\partial K}) - \Pi_{\partial K}^k(\delta_{|\partial K}).$$

Invoking the triangle inequality, the L^2-stability of $\Pi_{\partial K}^k$, and a discrete trace inequality leads to

$$\|\Pi_{\partial K}^k(\delta_{|\partial K})\|_{L^2(\partial K)} \le \|\mathsf{S}(\hat{\mathcal{I}}_K^k(v))\|_{L^2(\partial K)} + c\, h_K^{-\frac{1}{2}} \|\Pi_K^k(\delta)\|_{L^2(K)}.$$

Hence, we have

$$\|\nabla\delta\|_{\boldsymbol{L}^2(K)} \le c\left(h_K^{-1}\|\Pi_K^k(\delta)\|_{L^2(K)} + h_K^{-\frac{1}{2}}\|\mathsf{S}(\hat{\mathcal{I}}_K^k(v))\|_{L^2(\partial K)}\right),$$

and the assertion follows by invoking the bound from Step (i).
(iii) The Pythagorean identity implies that

$$\|\nabla v\|_{\boldsymbol{L}^2(K)}^2 = \|\nabla\mathcal{E}_K(v)\|_{\boldsymbol{L}^2(K)}^2 + \|\nabla(v-\mathcal{E}_K(v))\|_{\boldsymbol{L}^2(K)}^2.$$

The bound from Step (ii) implies that

$$\|\nabla(v-\mathcal{E}_K(v))\|_{\boldsymbol{L}^2(K)}^2 \le c\,\mathfrak{a}_K(v,v).$$

Since $\|\nabla\mathcal{E}_K(v)\|_{\boldsymbol{L}^2(K)}^2 \le \mathfrak{a}_K(v,v)$, this completes the proof.

Exercise 39.3 (Elliptic projection). Let us define the function $\delta \in H^1(D)$ such that $\delta_{|K} := u_{|K} - \mathcal{E}_K(u)$ on all $K \in \mathcal{T}_h$. Recall that $u \in H^{1+r}(D)$, $r > \frac{1}{2}$, and that $t := \min(k+1, r)$. Owing to the optimality property of the local elliptic projection and the approximation properties of Π_K^{k+1}, we have

$$\|\nabla\delta\|_{\boldsymbol{L}^2(K)} \le \|\nabla(u-\Pi_K^{k+1}(u))\|_{\boldsymbol{L}^2(K)} \le c\, h_K^t |u|_{H^{1+t}(K)}.$$

Using the triangle inequality, the approximation properties of Π_K^{k+1}, a discrete trace inequality, and the optimality property of the local elliptic projection, we infer that

$$\begin{aligned}
h_K^{\frac{1}{2}}\|\nabla\delta\|_{\boldsymbol{L}^2(\partial K)} &\le h_K^{\frac{1}{2}}\|\nabla(u-\Pi_K^{k+1}(u))\|_{\boldsymbol{L}^2(\partial K)} + h_K^{\frac{1}{2}}\|\nabla(\mathcal{E}_K(u)-\Pi_K^{k+1}(u))\|_{\boldsymbol{L}^2(\partial K)} \\
&\le c\left(h_K^t|u|_{H^{1+t}(K)} + \|\nabla(\mathcal{E}_K(u)-\Pi_K^{k+1}(u))\|_{\boldsymbol{L}^2(K)}\right) \\
&\le c\left(h_K^t|u|_{H^{1+t}(K)} + 2\|\nabla(u-\Pi_K^{k+1}(u))\|_{\boldsymbol{L}^2(K)}\right) \\
&\le c'\, h_K^t |u|_{H^{1+t}(K)}.
\end{aligned}$$

Exercise 39.4 (Reconstruction). (i) For all $q \in V_K^{k+1}$, since $\nabla V_K^{k+1} \subsetneq \boldsymbol{V}_K^k$, we have

$$\begin{aligned}
(\mathsf{G}(\hat{v}_K), \nabla q)_{\boldsymbol{L}^2(K)} &= -(v_K, \nabla\cdot\nabla q)_{L^2(K)} + (v_{\partial K}, \boldsymbol{n}_K\cdot\nabla q)_{L^2(\partial K)} \\
&= (\nabla\mathsf{R}(\hat{v}_K), \nabla q)_{\boldsymbol{L}^2(K)}.
\end{aligned}$$

This proves that $\Pi_{\nabla V_K^{k+1}}\mathsf{G} = \nabla\mathsf{R}$.
(ii) Let $\hat{v}_K \in \hat{V}_K^k$. Recalling from Definition 14.10 the dofs of the Raviart–Thomas finite element, we consider the function $\boldsymbol{q}_v \in \boldsymbol{V}_K^k$ s.t.

$$\begin{aligned}
(\boldsymbol{q}_v\cdot\boldsymbol{\nu}_K, \zeta_m\circ\boldsymbol{T}_F^{-1})_{L^2(F)} &= h_K^{-1}(v_{\partial K}-v_K, \zeta_m\circ\boldsymbol{T}_F^{-1})_{L^2(F)}, \quad \forall F\in\mathcal{F}_K, \\
(\boldsymbol{q}_v, \boldsymbol{\psi}_m\circ\boldsymbol{T}_K^{-1})_{\boldsymbol{L}^2(K)} &= (\nabla v_K, \boldsymbol{\psi}_m\circ\boldsymbol{T}_K^{-1})_{\boldsymbol{L}^2(K)},
\end{aligned}$$

where $\{\zeta_m\}_{m\in\{1:n_{\mathrm{sh}}^{\mathrm{f}}\}}$ is a basis of $\mathbb{P}_{k,d-1}$ with $n_{\mathrm{sh}}^{\mathrm{f}} := \dim(\mathbb{P}_{k,d-1}) = \binom{d+k-1}{k}$ and $\{\boldsymbol{\psi}_m\}_{m\in\{1:n_{\mathrm{sh}}^{\mathrm{c}}\}}$ is a basis of $\boldsymbol{\mathbb{P}}_{k-1,d}$ with $n_{\mathrm{sh}}^{\mathrm{c}} := \dim(\mathbb{P}_{k-1,d}) = \binom{d+k-1}{k-1}$. We observe that

$$\begin{aligned}
(\mathsf{G}_{\mathrm{RT}}(\hat{v}_K), \boldsymbol{q}_v)_{\boldsymbol{L}^2(K)} &= -(v_K, \nabla\cdot\boldsymbol{q}_v)_{L^2(K)} + (v_{\partial K}, \boldsymbol{n}_K\cdot\boldsymbol{q}_v)_{L^2(\partial K)} \\
&= (\nabla v_K, \boldsymbol{q}_v)_{\boldsymbol{L}^2(K)} - (v_K - v_{\partial K}, \boldsymbol{n}_K\cdot\boldsymbol{q}_v)_{L^2(\partial K)} \\
&= \|\nabla v_K\|_{\boldsymbol{L}^2(K)}^2 + h_K^{-1}\|v_K - v_{\partial K}\|_{L^2(\partial K)}^2 = |\hat{v}_K|_{\hat{V}_K^k}^2.
\end{aligned}$$

Using inverse inequalities shows that

$$\|\boldsymbol{q}_v\|_{\boldsymbol{L}^2(K)} \le c\left(\|\nabla v_K\|_{\boldsymbol{L}^2(K)} + h_K^{-1}\|v_K - v_{\partial K}\|_{L^2(\partial K)}\right) = c\,|\hat{v}_K|_{\hat{V}_K^k}.$$

The Cauchy–Schwarz inequality implies that

$$\begin{aligned}|\hat{v}_K|_{\hat{V}_K^k}^2 = (\mathsf{G}_{\mathrm{RT}}(\hat{v}_K), \boldsymbol{q}_v)_{\boldsymbol{L}^2(K)} &\le \|\mathsf{G}_{\mathrm{RT}}(\hat{v}_K)\|_{\boldsymbol{L}^2(K)}\|\boldsymbol{q}_v\|_{\boldsymbol{L}^2(K)}\\ &\le c\,\|\mathsf{G}_{\mathrm{RT}}(\hat{v}_K)\|_{\boldsymbol{L}^2(K)}|\hat{v}_K|_{\hat{V}_K^k},\end{aligned}$$

whence the conclusion.

Exercise 39.5 ($k = 0$). (i) We enumerate the mesh vertices from 0 to $N+1$. On a mesh cell $K_i := [x_i, x_{i+1}]$ of size h_i, for all $i \in \{0{:}N\}$, the discrete unknowns are the real number $u_i := u_{K_i}$ attached to the cell and the two real numbers $u_{\partial K_i} := (\lambda_i, \lambda_{i+1})$ attached to the two endpoints. Recall that $\lambda_0 = \lambda_{N+1} = 0$ owing to the enforcement of the homogeneous Dirichlet condition. A direct computation shows that $\frac{d}{dx}\mathsf{R}(u_i, (\lambda_i, \lambda_{i+1})) = h_i^{-1}(\lambda_{i+1} - \lambda_i)$ and that $\mathsf{S}(u_i, (\lambda_i, \lambda_{i+1})) = u_i - \frac{1}{2}(\lambda_i + \lambda_{i+1})$ at both endpoints of K_i. The local discrete equations are for all $v := (v_i)_{i\in\{0:N\}}$ and all $\mu := (\mu_i)_{i\in\{0:N+1\}}$ with $\mu_0 = \mu_{N+1} = 0$,

$$\sum_{i\in\{0:N\}}\Bigg(h_i^{-1}(\lambda_{i+1} - \lambda_i)(\mu_{i+1} - \mu_i) + 2h_i^{-1}\Big(u_i - \frac{1}{2}(\lambda_i + \lambda_{i+1})\Big)\Big(v_i - \frac{1}{2}(\mu_i + \mu_{i+1})\Big)\Bigg) = \sum_{i\in\{0:N\}} h_i\bar{f}_i v_i,$$

where $\bar{f}_i$ denotes the mean value of f over K_i. Taking first $v_i = \frac{1}{2}(\mu_i + \mu_{i+1})$ for all $i \in \{0{:}N\}$ allows us to get rid of the stabilization term and leads to

$$h_i^{-1}(\lambda_{i+1} - \lambda_i)(\mu_{i+1} - \mu_i) = h_i\bar{f}_i\frac{1}{2}(\mu_i + \mu_{i+1}).$$

This is the transmission problem identified in Proposition 39.10. The algebraic realization of this problem is $\mathcal{A}\Lambda = \mathsf{F}$, where $\mathcal{A}$ is the tridiagonal matrix of order N with entries $(-1, 2, -1)$, $\Lambda \in \mathbb{R}^N$ is the vector formed by the λ_i's at the interior vertices, and $\mathsf{F} \in \mathbb{R}^N$ has components given by $\mathsf{F}_i := \frac{1}{2}(h_{i-1}^2\bar{f}_{i-1} + h_i^2\bar{f}_i)$ for all $i \in \{1{:}N\}$. (Up to an aproximation of the right-hand side with a quadrature, this linear system is the same as the one obtained using continuous $\mathbb{P}_1$ Lagrange finite elements.) Once the λ_i's have been computed, the cell unknowns u_i are recovered by taking arbitrary cell test functions and zero face test functions. This gives

$$u_i = \frac{1}{2}h_i^2\bar{f}_i + \frac{1}{2}(\lambda_i + \lambda_{i+1}), \qquad \forall i \in \{0{:}N\}.$$

Finally, the local liftings are such that

$$U_{\lambda_i,\lambda_{i+1}} = \frac{1}{2}(\lambda_i + \lambda_{i+1}), \qquad U_{f_{|K}} = \frac{1}{2}h_i^2\bar{f}_i.$$

One can observe that $\frac{d}{dx}\mathsf{R}(U_{\lambda_i,\lambda_{i+1}}, (\lambda_i, \lambda_{i+1})) = h_i^{-1}(\lambda_{i+1} - \lambda_i)$ and that $\mathsf{S}(U_{\lambda_i,\lambda_{i+1}}, (\lambda_i, \lambda_{i+1})) = 0$. (ii) Let $\hat{v}_K \in \hat{V}_K^0$. Thus, v_K is constant on K and $v_{\partial K}$ is piecewise constant on ∂K. Let us set $v_F := v_{\partial K|F}$ for all $F \in \mathcal{F}_K$. For all $q \in \mathbb{P}_{1,d} \circ \boldsymbol{T}_K^{-1}$ with $q(\boldsymbol{x}) := q_K + \boldsymbol{G}_q\cdot(\boldsymbol{x} - \boldsymbol{x}_K)$ where $\boldsymbol{x}_K$ is the barycenter of K, we have

$$\begin{aligned}(\nabla\mathsf{R}(\hat{v}_K), \nabla q)_{\boldsymbol{L}^2(K)} &= -(v_K, \Delta q)_{L^2(K)} + (v_{\partial K}, \boldsymbol{n}_K\cdot\nabla q)\\ &= \sum_{F\in\mathcal{F}_K}|F|v_F\boldsymbol{n}_{K|F}\cdot\boldsymbol{G}_q.\end{aligned}$$

Since $\nabla\mathsf{R}(\hat{v}_K)$ and $\boldsymbol{G}_q$ are constant in K, we conclude that

$$\nabla\mathsf{R}(\hat{v}_K) = \sum_{F\in\mathcal{F}_K} \frac{|F|}{|K|} v_F \boldsymbol{n}_{K|F}.$$

Since $\sum_{F\in\mathcal{F}_K} |F|\boldsymbol{n}_{K|F} = \boldsymbol{0}$ (see (7.1)), we infer that

$$\nabla\mathsf{R}(\hat{v}_K) = \sum_{F\in\mathcal{F}_K} \frac{|F|}{|K|} (v_F - v_K) \boldsymbol{n}_{K|F},$$

and, finally, we obtain

$$\mathsf{R}(\hat{v}_K)(\boldsymbol{x}) = v_K + \sum_{F\in\mathcal{F}_K} \frac{|F|}{|K|} (v_F - v_K) \boldsymbol{n}_{K|F}\cdot(\boldsymbol{x} - \boldsymbol{x}_K), \qquad \forall \boldsymbol{x}\in K.$$

Furthermore, we have for all $F\in\mathcal{F}_K$,

$$\mathsf{S}(\hat{v}_K)_{|F} = \Pi_F^0(v_K - v_F + (I - \Pi_k^0)\mathsf{R}(\hat{v}_K)) = v_K - v_F - \nabla\mathsf{R}(\hat{v}_K)\cdot(\boldsymbol{x}_K - \boldsymbol{x}_F),$$

since $\Pi_K^0(\mathsf{R}(\hat{v}_K)) = v_K$ and $\Pi_F^0(\mathsf{R}(\hat{v}_K)) = v_K + \nabla\mathsf{R}(\hat{v}_K)\cdot(\boldsymbol{x}_F - \boldsymbol{x}_K)$.

Exercise 39.6 (Transmission problem). (i) Let $\hat{w}_h \in \hat{V}_{h,0}^k$ and assume that $u_{\mathcal{F}_h}$ solves the transmission problem (39.20). Setting $\hat{u}_K := (u_K, u_{\partial K}) := (U_{f_{|K}}, 0) + (U_{u_{\partial K}}, u_{\partial K})$ for all $K\in\mathcal{T}_h$, we infer that

$$\begin{aligned}
&\hat{a}_K(\hat{u}_K, \hat{w}_K)\\
&= \hat{a}_K((U_{f_{|K}},0) + (U_{u_{\partial K}}, u_{\partial K}), (w_K - U_{w_{\partial K}}, 0)) + \hat{a}_K((U_{f_{|K}},0) + (U_{u_{\partial K}}, u_{\partial K}), (U_{w_{\partial K}}, w_{\partial K}))\\
&= \hat{a}_K((U_{f_{|K}},0), (w_K - U_{w_{\partial K}}, 0)) + \ell_K(U_{w_{\partial K}}) + \hat{a}_K((U_{u_{\partial K}}, u_{\partial K}), (U_{w_{\partial K}}, w_{\partial K})) - \ell_K(U_{w_{\partial K}})\\
&= \ell_K(w_K) + \hat{a}_K((U_{u_{\partial K}}, u_{\partial K}), (U_{w_{\partial K}}, w_{\partial K})) - \ell_K(U_{w_{\partial K}}),
\end{aligned}$$

using that $\hat{a}_K((U_{u_{\partial K}}, u_{\partial K}), (y_K, 0)) = 0$ for all $y_K \in V_K^k$, a similar argument for $(U_{w_{\partial K}}, w_{\partial K})$ together with the symmetry of $\hat{a}_K$, and the definition of $U_{f_{|K}}$. Summing over $K\in\mathcal{T}_h$ shows that

$$\sum_{K\in\mathcal{T}_h} \hat{a}_K(\hat{u}_K, \hat{w}_K) = \sum_{K\in\mathcal{T}_h} \ell_K(w_K),$$

i.e., $\hat{u}_h$ solves the global HHO problem (39.16).
(ii) Let us prove the forward statement. Assume that u is the weak solution, i.e., $u\in H_0^1(D)$ and $a(u,w) = \ell(w)$ for all $w\in H_0^1(D)$. Let $K\in\mathcal{T}_h$. Let us now define λ by setting $\lambda_{\partial K} := u_{|\partial K}$ for all $K\in\mathcal{T}_h$. This definition makes sense, i.e., $\lambda_{\partial K}$ is single-valued since u does not jump across the interfaces (see Theorem 18.8). Moreover, $\lambda\in\Lambda$ owing to the trace theorem (see Theorem 3.10). Let $\psi\in H_0^1(K)$ and let $\tilde{\psi}$ be the zero-extension of ψ to D. Since $\tilde{\psi}\in H_0^1(D)$, we infer that $a(u,\tilde{\psi}) = \ell(\tilde{\psi})$. The definition of $U_{f_{|K}}$ implies that

$$a_K(u - U_{f_{|K}}, \psi) = a(u,\tilde{\psi}) - a_K(U_{f_{|K}}, \psi) = \ell(\tilde{\psi}) - \ell_K(\psi) = 0.$$

Since $(u - U_{f_{|K}})_{|\partial K} = \lambda_{\partial K}$, the above identity together with the definition of $U_{\lambda_{\partial K}}$ implies that $u_{|K} - U_{f_{|K}} = U_{\lambda_{\partial K}}$. Let us now prove that (39.22) holds true. Let μ be a member of Λ. Using the symmetry of a_K, the fact that $u_{|K} - U_{\lambda_{\partial K}} \in H_0^1(K)$, and the definition of $U_{\mu_{\partial K}}$, we infer that

$$a_K(u - U_{\lambda_{\partial K}}, U_{\mu_{\partial K}}) = a_K(U_{\mu_{\partial K}}, u - U_{\lambda_{\partial K}}) = 0.$$

As a result, we have $a_K(u,U_{\mu_{\partial K}})=a_K(U_{\lambda_{\partial K}},U_{\mu_{\partial K}})$. Consider the function w defined by setting $w_{|K}:=U_{\mu_{\partial K}}$ for all $K\in\mathcal{T}_h$. The restriction of w to K is in $H^1(K)$ for every $K\in\mathcal{T}_h$, the jumps of w across the interfaces vanish by construction, and w vanishes at the boundary faces. Theorem 18.8 implies that $w\in H^1_0(D)$. Using that $0=a(u,w)-\ell(w)$, this implies that

$$0=\sum_{K\in\mathcal{T}_h}\big(a_K(u,U_{\mu_{\partial K}})-\ell_K(U_{\mu_{\partial K}})\big)=\sum_{K\in\mathcal{T}_h}\big(a_K(U_{\lambda_{\partial K}},U_{\mu_{\partial K}})-\ell_K(U_{\mu_{\partial K}})\big),$$

thereby showing that $U_{\lambda_{\partial K}}$ solves the global transmission problem (39.22).

Let us now prove the converse statement. Assume that λ solves (39.22). Set $u:=U_{\lambda_{\partial K}}+U_{f_{|K}}$ for all $K\in\mathcal{T}_h$. This implies that $u_{|\partial K}=\lambda_{\partial K}$. Let $w\in H^1_0(D)$. Let $K\in\mathcal{T}_h$ and set $\mu:=w_{\partial K}$. Since $w-U_\mu\in H^1_0(K)$, we infer that

$$\begin{aligned}a_K(u,w-U_\mu)-\ell_K(w-U_\mu)&=a_K(U_{\lambda_{\partial K}},w-U_\mu)+a_K(U_{f_{|K}},w-U_\mu)-\ell_K(w-U_\mu)\\&=0+0=0,\end{aligned}$$

by definition of $U_{\lambda_{\partial K}}$ and $U_{f_{|K}}$. As a result, we have

$$\begin{aligned}a_K(u,w)-\ell_K(w)&=a_K(u,U_\mu)-\ell_K(U_\mu)+a_K(u,w-U_\mu)-\ell_K(w-U_\mu)\\&=a_K(u,U_\mu)-\ell_K(U_\mu)\\&=a_K(U_{f_{|K}},U_\mu)+a_K(U_{\lambda_{\partial K}},U_\mu)-\ell_K(U_\mu)\\&=a_K(U_{\lambda_{\partial K}},U_\mu)-\ell_K(U_\mu),\end{aligned}$$

since $a_K(U_{f_{|K}},U_\mu)=a_K(U_\mu,U_{f_{|K}})$ owing to the symmetry of a_K, the definition of $U_{\mu_{\partial K}}$, and the fact that $U_{f_{|K}}\in H^1_0(K)$. Summing the above equality for all $K\in\mathcal{T}_h$, we infer that u is the weak solution.

(iii) The global transmission problem (39.22) now consists of seeking $\lambda\in\Lambda$ such that

$$\sum_{K\in\mathcal{T}_h}a_K(U_{\lambda_{\partial K}},U^*_{\mu_{\partial K}})=\sum_{K\in\mathcal{T}_h}\ell_K(U^*_{\mu_{\partial K}}),\qquad\forall\mu\in\Lambda.$$

All the above arguments are readily adapted to this case.

(iv) Using that $U_{f_{|K}}\in H^1_0(K)$, $-\Delta U_{f_{|K}}=f_{|K}$, and $-\Delta U_{\lambda_{\partial K}}=0$, we have

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}\big(a_K(U_{\lambda_{\partial K}},U_{\mu_{\partial K}})-\ell_K(U_{\mu_{\partial K}})\big)\\&=\sum_{K\in\mathcal{T}_h}\big((\nabla U_\lambda,\nabla U_{\mu_{\partial K}})_{\boldsymbol{L}^2(K)}-(f,U_{\mu_{\partial K}})_{L^2(K)}\big)\\&=\sum_{K\in\mathcal{T}_h}\big((\boldsymbol{n}_K\cdot\nabla U_\lambda,U_{\mu_{\partial K}})_{L^2(\partial K)}+(\Delta U_f,U_{\mu_{\partial K}})_{L^2(K)}\big)\\&=\sum_{K\in\mathcal{T}_h}\big((\boldsymbol{n}_K\cdot\nabla(U_\lambda+U_f),U_{\mu_{\partial K}})_{L^2(\partial K)}-(\nabla U_f,\nabla U_{\mu_{\partial K}})_{\boldsymbol{L}^2(K)}\big)\\&=\sum_{K\in\mathcal{T}_h}(\nabla u\cdot\boldsymbol{n}_K,\mu)_{L^2(\partial K)}=\sum_{F\in\mathcal{F}_h^\circ}(\llbracket\nabla u\rrbracket_F\cdot\boldsymbol{n}_F,\mu)_{L^2(F)}.\end{aligned}$$

Exercise 39.7 (HDG). Since (39.25) amounts to a square finite-dimensional linear system, it suffices to prove that the only solution corresponding to zero data is the trivial one. Let $(\boldsymbol{\sigma}_{\mathcal{T}_h},u_{\mathcal{T}_h},\lambda_{\mathcal{F}_h})$

be one such solution. For all $K \in \mathcal{T}_h$, testing with $\boldsymbol{\tau}_K := \boldsymbol{\sigma}_K := \boldsymbol{\sigma}_{\mathcal{T}_h|K}$ and $w_K := u_K := u_{\mathcal{T}_h|K}$, and letting $\lambda_{\partial K} := (\lambda_F)_{F\in\mathcal{F}_K}$, we infer that

$$(\boldsymbol{\sigma}_K, \boldsymbol{\sigma}_K)_{\boldsymbol{L}^2(K)} + (\tau_{\partial K}(u_{K|\partial K} - \lambda_{\partial K}), u_{K|\partial K} - \lambda_{\partial K})_{L^2(\partial K)} = 0,$$

where we integrated by parts and used the expression (39.26) for the numerical flux trace $\boldsymbol{\phi}_h$. This implies that the two terms on the left-hand side vanish. In particular, we obtain $\boldsymbol{\sigma}_K = \mathbf{0}$. Equation (39.25a) with $\boldsymbol{\sigma}_K := \mathbf{0}$ and the fact that $(\tau_{\partial K}(u_{K|\partial K} - \lambda_{\partial K}), u_{K|\partial K} - \lambda_{\partial K})_{L^2(\partial K)} = 0$ imply, by assumption, that u_K and $\lambda_{\partial K}$ are constant functions taking the same value. Reasoning as we did in the argumentation above Lemma 39.8 and observing that $\lambda_{\mathcal{F}_h|\mathcal{F}_h^\partial} = 0$ by construction, we conclude that $u_{\mathcal{T}_h}$ and $\lambda_{\mathcal{F}_h}$ vanish.

Exercise 39.8 (Space Λ). (i) Let $\widehat{K}$ be the reference cell. Let $\widehat{\mu} \in H^{\frac12}(\partial\widehat{K})$. Since the trace map $\gamma^{\mathrm{g}}_{\partial\widehat{K}}$ is surjective, we have

$$\widehat{c}_1 |\widehat{\mu}|_{H^{\frac12}(\partial\widehat{K})} \le \inf_{\substack{\widehat{v}\in H^1(\widehat{K})\\ \gamma^{\mathrm{g}}_{\partial\widehat{K}}(\widehat{v})=\widehat{\mu}}} \|\nabla\widehat{v}\|^2_{\boldsymbol{L}^2(\widehat{K})} \le \|\nabla\widehat{U}_{\widehat{\mu}}\|^2_{\boldsymbol{L}^2(\widehat{K})},$$

where $\widehat{U}_{\widehat{\mu}}$ is the unique solution in $H^1(\widehat{K})$ s.t.

$$\gamma^{\mathrm{g}}_{\partial\widehat{K}}(\widehat{U}_{\widehat{\mu}}) = \widehat{\mu}, \qquad (\nabla\widehat{U}_{\widehat{\mu}}, \nabla\widehat{\psi})_{\boldsymbol{L}^2(\widehat{K})} = 0,\ \forall\widehat{\psi} \in H^1_0(\widehat{K}).$$

Reasoning as in Proposition 31.12, i.e., invoking the stability of the Poisson problem with non-homogeneous Dirichlet conditions, yields

$$\|\nabla\widehat{U}_{\widehat{\mu}}\|_{\boldsymbol{L}^2(K)} \le \widehat{c}_2\, |\widehat{\mu}|_{H^{\frac12}(\partial\widehat{K})}.$$

Let now $K \in \mathcal{T}_h$ and $\mu \in H^{\frac12}(\partial K)$. Let us set $\widehat{\mu} := \mu\circ\boldsymbol{T}_{K|\partial\widehat{K}}$. Using the transformation of Sobolev seminorms by the pullback by the geometric mapping (see Lemma 11.7), we infer that

$$|\mu|_{H^{\frac12}(\partial K)} \le \widehat{c}\, h_K^{\frac{d-1}{2}-\frac12} |\widehat{\mu}|_{H^{\frac12}(\partial\widehat{K})} \le \widehat{c}\widehat{c}_1^{-1}\, h_K^{\frac{d}{2}-1} \|\nabla\widehat{U}_{\widehat{\mu}}\|_{\boldsymbol{L}^2(\widehat{K})}.$$

Moreover, since $U_\mu\circ\boldsymbol{T}_K \in H^1(\widehat{K})$ and $\gamma^{\mathrm{g}}_{\partial\widehat{K}}(U_\mu\circ\boldsymbol{T}_K) = \widehat{\mu}$, we infer using (9.8a) that

$$\begin{aligned}\|\nabla\widehat{U}_{\widehat{\mu}}\|_{\boldsymbol{L}^2(\widehat{K})} &\le \|\nabla(U_\mu\circ\boldsymbol{T}_K)\|_{\boldsymbol{L}^2(\widehat{K})} = \|\mathbb{J}_K^{\mathsf{T}}(\nabla U_\mu)\circ\boldsymbol{T}_K\|_{\boldsymbol{L}^2(\widehat{K})}\\ &\le \frac{|\widehat{K}|^{\frac12}}{|K|^{\frac12}} \|\mathbb{J}_K^{\mathsf{T}}\|_{\ell^2} \|\nabla U_\mu\|_{\boldsymbol{L}^2(K)} \le \widehat{c}\, h_K^{1-\frac{d}{2}} \|\nabla U_\mu\|_{\boldsymbol{L}^2(K)}.\end{aligned}$$

Hence, we have $|\mu|_{H^{\frac12}(\partial K)} \le c\|\nabla U_\mu\|_{\boldsymbol{L}^2(K)}$, where c only depends on the regularity of the mesh sequence. The proof of the converse bound uses similar arguments.
(ii) The only nontrivial property to verify in order to prove that $\|\lambda\|^2_\Lambda := \sum_{K\in\mathcal{T}_h} |\lambda_{\partial K}|^2_{H^{\frac12}(\partial K)}$ defines a norm on Λ is that $\|\lambda\|_\Lambda = 0$ implies that $\lambda = 0$. Assuming that $\|\lambda\|_\Lambda = 0$, we infer that $\lambda_{\partial K}$ is constant for all $K \in \mathcal{T}_h$. Since $\lambda_{|\mathcal{F}_h^\partial} = 0$, this implies that $\lambda_{\partial K}$ is zero for all $K \in \mathcal{T}_h$ having at least one boundary face. We can then repeat the argument for cells having an interface with those cells, and we can move inward and reach all the cells in $\mathcal{T}_h$ by repeating the process a finite number of times. This proves that $\lambda = 0$.

For all $\lambda \in \Lambda$, let $U_\lambda : D \to \mathbb{R}$ be s.t. $U_{\lambda|K} := U_{\lambda_{\partial K}}$ for all $K \in \mathcal{T}_h$, that is, we have $U_{\lambda|K} \in H^1(K)$, $a_K(U_{\lambda|K}, \phi) = 0$ for all $\phi \in H^1_0(K)$, and $\gamma^{\mathrm{g}}_{\partial K}(U_\lambda) = \lambda_{\partial K}$ for all $K \in \mathcal{T}_h$. By

definition of the space Λ, the lifting U_λ vanishes on all the boundary faces and is continuous across all the mesh interfaces. Theorem 18.8 then implies that $U_\lambda \in H^1_0(D)$. Consider now a Cauchy sequence $(\lambda_n)_{n\in\mathbb{N}}$ in Λ, and let us set $U_n := U_{\lambda_n} \in H^1_0(D)$ for all $n \in \mathbb{N}$. Summing the result from Step (i) over all the mesh cells, we infer that

$$c_1 \|U_n\|_{H^1_0(D)} \le \|\lambda_n\|_\Lambda \le c_2 \|U_n\|_{H^1_0(D)}, \tag{39.1}$$

for all $n \in \mathbb{N}$, with $0 < c_1 \le c_2$, where we equipped the Hilbert space $H^1_0(D)$ with the norm $\|v\|_{H^1_0(D)} := \|\nabla v\|_{\boldsymbol{L}^2(D)}$ owing to the Poincaré–Steklov inequality. The lower bound in (39.1) implies that $(U_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $H^1_0(D)$. Hence, there is $U \in H^1_0(D)$ s.t. $U_n \to U$ in $H^1_0(D)$ as $n \to \infty$. The function $\lambda \in L^2(\mathcal{F}_h)$ s.t. $\lambda_{\partial K} := \gamma^{\rm g}_{\partial K}(U)$ is in Λ since U has a zero trace on the boundary faces. Moreover, the upper bound in (39.1) implies that $\lambda_n \to \lambda$ in Λ as $n \to \infty$. This proves that Λ equipped with the norm $\|\cdot\|_\Lambda$ is a Hilbert space.

Exercise 39.9 (Liftings, 1D). (i) Let $K_i := [x_{i-1}, x_i]$ be a mesh cell. In 1D, we identify $H^{\frac12}(\partial K_i)$ with $\mathbb{R}^2$, and we write $\lambda := (\lambda_{i-1}, \lambda_i)^\mathsf{T}$. Since $(U_\lambda)'' = 0$ in K_i, we infer that U_λ is affine in K. Since $U_\lambda(x_{i-1}) = \lambda_{i-1}$ and $U_\lambda(x_i) = \lambda_i$, we infer that $U_\lambda = (\lambda_{i-1}\varphi_{i-1} + \lambda_i\varphi_i)_{|K_i}$, where the φ_i's are the global shape functions in 1D. Inserting this expression into (39.22), we obtain the linear system $\mathcal{A}X = B$ with $X_i := \lambda_i$. It remains to prove that $\lambda_i = u(x_i)$ for all $i \in \{1{:}I\}$. Since $f = -u''$, integration by parts leads to

$$B_i = -\int_{x_{i-1}}^{x_{i+1}} u''\varphi_i \,\mathrm{d}s = \int_{x_{i-1}}^{x_{i+1}} u'\varphi_i' \,\mathrm{d}s = \frac{1}{h}\int_{x_{i-1}}^{x_i} u' \,\mathrm{d}s - \frac{1}{h}\int_{x_i}^{x_{i+1}} u' \,\mathrm{d}s.$$

Hence, $B_i = -\frac{u(x_{i+1})-u(x_i)}{h} - \frac{u(x_i-u(x_{i-1})}{h}$. Using the matrix $\mathcal{A}$, we infer that $B = \mathcal{A}\bar{U}$, where $\bar{U}$ has components $\bar{U}_i = u(x_i)$. Hence, $\mathcal{A}(X - \bar{U}) = 0$, and since $\mathcal{A}$ is invertible, we infer that $X = \bar{U}$, i.e., $\lambda_i = u(x_i)$ for all $i \in \{1{:}I\}$.
(ii) The functions $\{\phi_m\}_{m\in\{1:k-1\}}$ are linearly independent. Moreover, ϕ_m is of degree $(m+1)$, and $\phi_m(\pm1) = 0$ since $L_m(-1) = (-1)^m$ and $L_m(1) = 1$. Hence, $\{\phi_m\}_{m\in\{1:k-1\}}$ is a basis of $\mathbb{P}^\circ_k$. Using the hint, we observe that $\int_K \phi_m'(x)\phi_l'(x)\,\mathrm{d}x = \delta_{kl}$. As a result, letting $\widehat{K} := [-1,1]$, we infer that for all $p \in \mathbb{P}^\circ_k$ with $p := \sum_{l\in\{1:k-1\}} p_l\phi_l$,

$$\begin{aligned}
\int_{\widehat{K}} (U_{f_{|\widehat{K}}})'(x)p'(x)\,\mathrm{d}x &= \sum_{m\in\{1:k-1\}} \int_{\widehat{K}}\int_{\widehat{K}} \phi_m'(x)\phi_m(s)f(s)p'(x)\,\mathrm{d}s\,\mathrm{d}x \\
&= \sum_{m\in\{1:k-1\}}\sum_{l\in\{1:k-1\}} p_l \int_{\widehat{K}}\int_{\widehat{K}} \phi_m'(x)\phi_m(s)f(s)\phi_l'(x)\,\mathrm{d}x\,\mathrm{d}s \\
&= \sum_{m\in\{1:k-1\}}\sum_{l\in\{1:k-1\}} p_l\delta_{lm} \int_{\widehat{K}} \phi_m(s)f(s)\,\mathrm{d}s = \int_{\widehat{K}} f(s)p(s)\,\mathrm{d}s.
\end{aligned}$$

This proves that $U_{f_{|\widehat{K}}}(x) = \int_{\widehat{K}} G(x,s)f(s)\,\mathrm{d}s$. The expression for $U_{f_{|K_i}}$ in K_i is obtained by applying the pullback by the map $\psi_i(t) = \frac{x_{i-1}+x_i}{2} + t\frac{h}{2}$ for all $t \in [-1,1]$. We obtain $U_{f_{|K_i}}(x) = \int_{K_i} G_i(x,s)f(s)\,\mathrm{d}s$ with $G_i(x,s) := \frac{x_i-x_{i-1}}{2}\sum_{m\in\{1:k-1\}}(\phi_m\circ\psi_i^{-1})(x)(\phi_m\circ\psi_i^{-1})(s)$.

Chapter 40

Contrasted diffusivity (I)

Exercises

Exercise 40.1 (Normal flux). Let $\boldsymbol{\sigma} \in \{\boldsymbol{\tau} \in \boldsymbol{L}^p(K) \mid \nabla\cdot\boldsymbol{\tau} \in L^2(K)\}$, $p > 2$. Let $\gamma_{\partial K}^{\mathrm{d}}(\boldsymbol{\sigma}) \in H^{-\frac12}(\partial K)$ be s.t. $\langle \gamma_{\partial K}^{\mathrm{d}}(\boldsymbol{\sigma}), \phi\rangle_{\partial K} := \int_K \boldsymbol{\sigma}\cdot\nabla v(\phi)\,\mathrm{d}x + \int_K (\nabla\cdot\boldsymbol{\sigma}) v(\phi)\,\mathrm{d}x$ for all $\phi \in H^{\frac12}(\partial K)$, where $v(\phi) \in H^1(K)$ is a lifting of ϕ, i.e., $\gamma_{\partial K}^{\mathrm{g}}(v(\phi)) = \phi$ (see (4.12)). Prove that $\langle \gamma_{\partial K}^{\mathrm{d}}(\boldsymbol{\sigma}), \phi\rangle_{\partial K} = \sum_{F\in\mathcal{F}_K} \langle (\boldsymbol{\sigma}\cdot\boldsymbol{n}_K)_{|F}, \phi_{|F}\rangle_F$. (*Hint*: reason as in the proof of (40.18b).)

Exercise 40.2 (Bound on $|v|_\sharp$). Prove that for all $v \in V_{\mathrm{s}}$, $|v|_{n_\sharp} \le c\lambda_\flat^{-\frac12}(\ell_D^{d(\frac12-\frac1p)}\|\boldsymbol{\sigma}(v)\|_{\boldsymbol{L}^p(D)} + \ell_D^{d(\frac{2+d}{2d}-\frac1q)}\|\nabla\cdot\boldsymbol{\sigma}(v)\|_{L^q(D)})$. (*Hint*: for the sum with $\boldsymbol{L}^p$-norms, use Hölder's inequality after observing that $h_K^d \le c|K|$, and for the sum with L^q norms, use that $(\sum_{K\in\mathcal{T}_h} a_K^t)^{\frac1t} \le (\sum_{K\in\mathcal{T}_h} a_K^s)^{\frac1s}$ for real numbers $t \ge s$.)

Exercise 40.3 (Jump identity). Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$. Let $\theta_l, \theta_r \in [0,1]$ be s.t. $\theta_l + \theta_r = 1$. Set $\{a\}_\theta := \theta_l a_l + \theta_r a_r$ and $\{a\}_{\bar\theta} := \theta_r a_l + \theta_l a_r$. (i) Show that $[\![ab]\!] = \{a\}_{\bar\theta}[\![b]\!] + [\![a]\!]\{b\}_\theta$. (ii) Show that $[\![ab]\!] = \{a\}_\theta[\![b]\!] + [\![a]\!]\{b\}_{\bar\theta}$.

Solution to exercises

Exercise 40.1 (Normal flux). Let $\phi \in H^{\frac12}(K)$ and let $v(\phi) \in H^1(K)$ s.t. $\gamma_{\partial K}^{\mathrm{g}}(v(\phi)) = \phi$. Consider the mollification operators $\mathcal{K}_\delta^{\mathrm{d}} : \boldsymbol{L}^1(D) \to \boldsymbol{C}^\infty(\overline{D})$ and $\mathcal{K}_\delta^{\mathrm{b}} : L^1(D) \to C^\infty(\overline{D})$ introduced in §23.1. Let us introduce the shorthand notation

$$\mathcal{F}_\delta(\phi) := \sum_{F\in\mathcal{F}_K} \langle (\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma})\cdot\boldsymbol{n}_K)_{|F}, \phi_{|F}\rangle_F.$$

Recalling (40.15), we have

$$\mathcal{F}_\delta(\phi) = \sum_{F\in\mathcal{F}_K} \int_K \big(\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma})\cdot\nabla L_F^K(\phi_{|F}) + \nabla\cdot(\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma})) L_F^K(\phi_{|F})\big)\,\mathrm{d}x,$$

and invoking the commuting property (40.19), we infer that

$$\mathcal{F}_\delta(\phi) = \sum_{F\in\mathcal{F}_K} \int_K \big(\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma})\cdot\nabla L_F^K(\phi_{|F}) + \mathcal{K}_\delta^{\mathrm{b}}(\nabla\cdot\boldsymbol{\sigma}))L_F^K(\phi_{|F})\big)\,\mathrm{d}x.$$

Therefore, we have

$$\begin{aligned}\lim_{\delta\to 0}\mathcal{F}_\delta(\phi) &= \sum_{F\in\mathcal{F}_K} \int_K \big(\boldsymbol{\sigma}\cdot\nabla L_F^K(\phi_{|F}) + (\nabla\cdot\boldsymbol{\sigma}))L_F^K(\phi_{|F})\big)\,\mathrm{d}x\\ &= \sum_{F\in\mathcal{F}_K} \langle(\boldsymbol{\sigma}\cdot\boldsymbol{n}_K)_{|F},\phi_{|F}\rangle_F.\end{aligned}$$

Since $\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma})$ is a smooth function, we also have

$$\begin{aligned}\mathcal{F}_\delta(\phi) &= \sum_{F\in\mathcal{F}_K} \int_{\partial K} \mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma})\cdot\boldsymbol{n}_K L_F^K(\phi_{|F})\,\mathrm{d}s\\ &= \int_{\partial K} \mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma})\cdot\boldsymbol{n}_K\phi\,\mathrm{d}s\\ &= \int_K \big(\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma})\cdot\nabla v(\phi) + \nabla\cdot(\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma}))v(\phi)\big)\,\mathrm{d}x\\ &= \int_K \big(\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma})\cdot\nabla v(\phi) + \mathcal{K}_\delta^{\mathrm{b}}(\nabla\cdot\boldsymbol{\sigma})v(\phi)\big)\,\mathrm{d}x.\end{aligned}$$

We infer that

$$\lim_{\delta\to 0}\mathcal{F}_\delta(\phi) = \int_K \big(\boldsymbol{\sigma}\cdot\nabla v(\phi) + (\nabla\cdot\boldsymbol{\sigma})v(\phi)\big)\,\mathrm{d}x,$$

and this concludes the proof.

Exercise 40.2 (Bound on $|v|_\sharp$). Let us write

$$\begin{aligned}\mathfrak{T}_1 &:= \Big(\sum_{K\in\mathcal{T}_h} h_K^{2d(\frac12-\frac1p)}\|\boldsymbol{\sigma}(v)_{|K}\|_{\boldsymbol{L}^p(K)}^2\Big)^{\frac12},\\ \mathfrak{T}_2 &:= \Big(\sum_{K\in\mathcal{T}_h} h_K^{2d(\frac{2+d}{2d}-\frac1q)}\|\nabla\cdot\boldsymbol{\sigma}(v)_{|K}\|_{L^q(K)}^2\Big)^{\frac12}.\end{aligned}$$

Concerning $\mathfrak{T}_1$, the regularity of the mesh sequence and Hölder's inequality with $r := \frac{p}{2} > 1$ and $r' := \frac{r}{r-1} = \frac{p}{p-2}$ imply that

$$\begin{aligned}\mathfrak{T}_1 &\le c\Big(\sum_{K\in\mathcal{T}_h} |K|^{2(\frac12-\frac1p)}\|\boldsymbol{\sigma}(v)_{|K}\|_{\boldsymbol{L}^p(K)}^2\Big)^{\frac12}\\ &= c\Big(\sum_{K\in\mathcal{T}_h} |K|^{\frac{1}{r'}}\|\boldsymbol{\sigma}(v)_{|K}\|_{\boldsymbol{L}^p(K)}^2\Big)^{\frac12}\\ &\le c\Big(\sum_{K\in\mathcal{T}_h} |K|\Big)^{\frac{1}{2r'}}\Big(\sum_{K\in\mathcal{T}_h} \|\boldsymbol{\sigma}(v)_{|K}\|_{\boldsymbol{L}^p(K)}^p\Big)^{\frac1p}\\ &= c\,|D|^{\frac12-\frac1p}\|\boldsymbol{\sigma}(v)\|_{\boldsymbol{L}^p(D)} \le c\,\ell_D^{d(\frac12-\frac1p)}\|\boldsymbol{\sigma}(v)\|_{\boldsymbol{L}^p(D)},\end{aligned}$$

where the last bound follows from $|D| \le \ell_D^d$. Moreover, since $(\sum_{K\in\mathcal{T}_h} a_K^t)^{\frac{1}{t}} \le (\sum_{K\in\mathcal{T}_h} a_K^s)^{\frac{1}{s}}$ for $t \ge s$, and since $q \le 2$, we infer that

$$\mathfrak{T}_2 \le \Big(\sum_{K\in\mathcal{T}_h} h_K^{qd(\frac{2+d}{2d}-\frac{1}{q})} \|\nabla\cdot\boldsymbol{\sigma}(v)_{|K}\|_{L^q(K)}^q \Big)^{\frac{1}{q}} \le \ell_D^{d(\frac{2+d}{2d}-\frac{1}{q})} \|\nabla\cdot\boldsymbol{\sigma}(v)\|_{L^q(D)},$$

where the last bound follows from $\frac{2+d}{2d} > \frac{1}{q}$ and $h_K \le h \le \ell_D$ for all $K \in \mathcal{T}_h$. In conclusion, we have shown that

$$|v|_{n_\sharp} \le c\,\lambda_\flat^{-\frac{1}{2}} \big(\ell_D^{d(\frac{1}{2}-\frac{1}{p})} \|\boldsymbol{\sigma}(v)\|_{\boldsymbol{L}^p(D)} + \ell_D^{d(\frac{2+d}{2d}-\frac{1}{q})} \|\nabla\cdot\boldsymbol{\sigma}(v)\|_{L^q(D)} \big).$$

Exercise 40.3 (Jump identity). (i) We verify the statement

$$\begin{aligned}
\{a\}_{\bar\theta}[\![b]\!] + [\![a]\!]\{b\}_\theta &= (\theta_r a_l + \theta_l a_r)(b_l - b_r) + (a_l - a_r)(\theta_l b_l + \theta_r b_r) \\
&= \theta_r a_l b_l - \theta_r a_l b_r + \theta_l a_r b_l - \theta_l a_r b_r + \theta_l a_l b_l + \theta_r a_l b_r - \theta_l a_r b_l - \theta_r a_r b_r \\
&= \theta_r a_l b_l - \theta_l a_r b_r + \theta_l a_l b_l - \theta_r a_r b_r \\
&= a_l b_l - a_r b_r = [\![ab]\!].
\end{aligned}$$

(ii) Switching a and b and using Step (i), we obtain

$$[\![ab]\!] = [\![ba]\!] = \{b\}_{\bar\theta}[\![a]\!] + [\![b]\!]\{a\}_\theta = \{a\}_\theta[\![b]\!] + [\![a]\!]\{b\}_{\bar\theta}.$$

Chapter 41

Contrasted diffusivity (II)

Exercises

Exercise 41.1 (Conforming finite elements). Consider the approximation of (40.3) by conforming finite elements. Let $V := H^1_0(D)$, $V_h := P^{\mathrm{g}}_{k,0}(\mathcal{T}_h) \subset V$, $k \geq 1$, and consider the norm $\|v\|_V := \|\lambda^{\frac12}\nabla v\|_{\boldsymbol{L}^2(D)}$. Assume $u \in H^{1+r}(D)$, $r > 0$, and set $t := \min(r, k)$. Prove that there is c, uniform w.r.t. λ, s.t. $\|u - u_h\|_V \leq c(\sum_{K\in\mathcal{T}_h} \lambda_K h_K^{2t}|u|^2_{H^{1+t}(\check{\mathcal{T}}_K)})^{\frac12}$ for all $h \in \mathcal{H}$, where $\check{\mathcal{T}}_K$ is the collection of the mesh cells sharing at least a vertex with K, and that $|u|_{H^{1+t}(\check{\mathcal{T}}_K)}$ can be replaced by $|u|_{H^{1+t}(K)}$ if $1 + t > \frac{d}{2}$.

Exercise 41.2 (dG). Prove the estimate (41.21).

Exercise 41.3 (HHO). (i) Prove (41.28a) (*Hint*: adapt the proof of (40.18a), i.e., use the definition of the pairing $\langle\cdot,\cdot\rangle_F$ together with the definition (39.2) for R). (ii) Prove (41.28b). (*Hint*: adapt the proof of (40.18b). (iii) Prove the error bound (41.31). (*Hint*: see the proof of (39.32) in Theorem 39.17.) (iv) Prove (41.32). (*Hint*: set $\ell = \lceil t\rceil$ and consider the elliptic projection of degree ℓ, say $\mathcal{E}^\ell_K$, for all $K \in \mathcal{T}_h$.)

Solution to exercises

Exercise 41.1 (Conforming finite elements). Reasoning as in the proof of Céa's lemma, we infer that

$$\|u - u_h\|_V = \inf_{v_h\in V_h} \|u - v_h\|_V.$$

We bound the infimum by taking $v_h := \mathcal{I}^{\mathrm{g,av}}_{h0}(u)$ of degree ℓ s.t. $\ell := \lceil t\rceil$. We can then invoke Theorem 22.14 to conclude that $\|u - u_h\|_V \leq c(\sum_{K\in\mathcal{T}_h} \lambda_K h_K^{2t}|u|^2_{H^{1+t}(\check{\mathcal{T}}_K)})^{\frac12}$. If $1 + t > \frac{d}{2}$, we can take instead $v_h := \mathcal{I}^{\mathrm{g}}_{h0}(u)$ (the canonical interpolation operator with zero boundary trace) or $v_h := \mathcal{I}^{\mathrm{L}}_{h0}(u)$ (the Lagrange interpolation operator with zero boundary trace). In both cases, we obtain that $\|u - u_h\|_V \leq c(\sum_{K\in\mathcal{T}_h} \lambda_K h_K^{2t}|u|^2_{H^{1+t}(K)})^{\frac12}$.

Exercise 41.2 (dG). The definition of $n_\sharp$ and the Cauchy–Schwarz inequality imply that

$$\begin{aligned}|n_\sharp(v_h,w_h)| &= \Big|\sum_{F\in\mathcal{F}_h}\int_F \{\boldsymbol{\sigma}(v_h)\}_\theta\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s\Big| \\ &\le \sum_{F\in\mathcal{F}_h} h_F^{\frac12}\lambda_F^{-\frac12}\|\{\boldsymbol{\sigma}(v_h)\}_\theta\cdot\boldsymbol{n}_F\|_{L^2(F)}\times\lambda_F^{\frac12}h_F^{-\frac12}\|[\![w_h]\!]\|_{L^2(F)} \\ &\le \sum_{F\in\mathcal{F}_h}\Big(\sum_{K\in\mathcal{T}_F}|\mathcal{T}_F|\theta_{K,F}^2\lambda_K\lambda_F^{-1}h_F\|\lambda_K^{\frac12}\nabla v_{h|K}\|_{\boldsymbol{L}^2(F)}^2\Big)^{\frac12}\times\lambda_F^{\frac12}h_F^{-\frac12}\|[\![w_h]\!]\|_{L^2(F)},\end{aligned}$$

where we used that

$$\begin{aligned}\|\{\boldsymbol{\sigma}(v_h)\}_\theta\cdot\boldsymbol{n}_F\|_{L^2(F)}^2 &= \int_F\Big|\sum_{K\in\mathcal{T}_F}\theta_{K,F}\lambda_K\nabla v_{h|K}\cdot\boldsymbol{n}_F\Big|^2\,\mathrm{d}s \\ &\le \sum_{K\in\mathcal{T}_F}|\mathcal{T}_F|\theta_{K,F}^2\lambda_K^2\int_F\|\nabla v_{h|K}\|_{\ell^2}^2\,\mathrm{d}s \\ &= \sum_{K\in\mathcal{T}_F}|\mathcal{T}_F|\theta_{K,F}^2\lambda_K\|\lambda_K^{\frac12}\nabla v_{h|K}\|_{\boldsymbol{L}^2(F)}^2.\end{aligned}$$

Using that $\theta_{K,F}\le\theta_{K,F}^{\frac12}$ (since $\theta_{K,F}\le1$) and that $|\mathcal{T}_F|\theta_{K,F}\lambda_K\lambda_F^{-1}=1$, together with the inverse inequality $h_F\|\nabla v_{h|K}\|_{\boldsymbol{L}^2(F)}^2\le c_{\mathrm{dt}}^2\|\nabla v_{h|K}\|_{\boldsymbol{L}^2(K)}^2$, and invoking the Cauchy–Schwarz inequality, we obtain

$$|n_\sharp(v_h,w_h)|\le c_{\mathrm{dt}}\Big(\sum_{F\in\mathcal{F}_h}\sum_{K\in\mathcal{T}_F}\|\lambda_K^{\frac12}\nabla v_{h|K}\|_{L^2(K)}^2\Big)^{\frac12}\times\Big(\sum_{F\in\mathcal{F}_h}\lambda_Fh_F^{-1}\|[\![w_h]\!]\|_{L^2(F)}^2\Big)^{\frac12}.$$

Since $\sum_{F\in\mathcal{F}_h}\sum_{K\in\mathcal{T}_F}(\cdot)=\sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K}(\cdot)$ and $\mathrm{card}(\mathcal{F}_K)\le n_\partial$, we conclude that

$$|n_\sharp(v_h,w_h)|\le c_{\mathrm{dt}}n_\partial^{\frac12}\|\lambda^{\frac12}\nabla_hv_h\|_{\boldsymbol{L}^2(D)}\Big(\sum_{F\in\mathcal{F}_h}\lambda_Fh_F^{-1}\|[\![w_h]\!]\|_{L^2(F)}^2\Big)^{\frac12}.$$

Exercise 41.3 (HHO). (i) Let $v_h\in P_{k+1}^{\mathrm{b}}(\mathcal{T}_h)$ and $\hat{w}_h\in\hat{V}_{h,0}^k$. Since the restriction of $\boldsymbol{\sigma}(v_h)$ to each mesh cell is smooth and since the trace on ∂K of the face-to-cell lifting operator L_F^K is nonzero only on F for all $F\in\mathcal{F}_K$, we have

$$\begin{aligned}\langle(\boldsymbol{\sigma}(v_h)\cdot\boldsymbol{n}_K)_{|F},(w_K-w_{\partial K})_{|F}\rangle_F &= \int_K\boldsymbol{\sigma}(v_h)_{|K}\cdot\nabla L_F^K((w_K-w_{\partial K})_{|F}) \\ &\quad+(\nabla\cdot\boldsymbol{\sigma}(v_h)_{|K})L_F^K((w_K-w_{\partial K})_{|F})\Big)\,\mathrm{d}x \\ &= \int_{\partial K}\boldsymbol{\sigma}(v_h)_{|K}\cdot\boldsymbol{n}_KL_F^K((w_K-w_{\partial K})_{|F})\,\mathrm{d}s \\ &= \int_F\boldsymbol{\sigma}(v_h)_{|K}\cdot\boldsymbol{n}_K(w_K-w_{\partial K})\,\mathrm{d}s,\end{aligned}$$

where we used the divergence formula in K. Therefore, we obtain

$$\begin{aligned} n_\sharp(v_h,\hat w_h) &= \sum_{K\in\mathcal T_h}\int_{\partial K}\boldsymbol\sigma(v_h)_{|K}\cdot\boldsymbol n_K(w_K-w_{\partial K})\,\mathrm ds\\ &= -\sum_{K\in\mathcal T_h}\lambda_K\int_{\partial K}\nabla v_{h|K}\cdot\boldsymbol n_K(w_K-w_{\partial K})\,\mathrm ds\\ &= \sum_{K\in\mathcal T_h}\lambda_K\int_K\big(\nabla v_{h|K}\cdot\nabla(\mathsf R(\hat w_K)-w_K)\big)\,\mathrm dx, \end{aligned}$$

where we used the definition (39.2) of the local reconstruction operator $\mathsf R$ with the test function $v_{h|K}\in\mathbb P_{k,d}\subset V_K^{k+1}$.

(ii) Let us now prove (41.28b). Let $v\in V_{\mathrm s}$ and $\hat w_h\in\hat V_{h,0}^k$. We are going to proceed as in the proof of (40.18b). We consider the mollification operators $\mathcal K_\delta^{\mathrm d}:\boldsymbol L^1(D)\to\boldsymbol C^\infty(\overline D)$ and $\mathcal K_\delta^{\mathrm b}:L^1(D)\to C^\infty(\overline D)$ introduced in §23.1. Let us consider the mollified bilinear form

$$n_{\sharp\delta}(v,\hat w_h):=\sum_{K\in\mathcal T_h}\sum_{F\in\mathcal F_K}\langle(\mathcal K_\delta^{\mathrm d}(\boldsymbol\sigma(v))\cdot\boldsymbol n_K)_{|F},(w_K-w_{\partial K})_{|F}\rangle_F.$$

By using (40.15) and invoking the approximation properties of the mollification operators and the commuting property (40.19), we infer that $\lim_{\delta\to0}n_{\sharp\delta}(v,\hat w_h)=n_\sharp(v,\hat w_h)$. Since the restriction of $\mathcal K_\delta^{\mathrm d}(\boldsymbol\sigma(v))$ to each mesh cell is smooth and since $\mathcal K_\delta^{\mathrm d}(\boldsymbol\sigma(v))\in\boldsymbol C^0(\overline D)$, we infer that

$$\begin{aligned} n_{\sharp\delta}(v,\hat w_h) &= \sum_{K\in\mathcal T_h}\int_{\partial K}\mathcal K_\delta^{\mathrm d}(\boldsymbol\sigma(v))\cdot\boldsymbol n_K(w_K-w_{\partial K})\,\mathrm ds\\ &= \sum_{K\in\mathcal T_h}\int_{\partial K}\mathcal K_\delta^{\mathrm d}(\boldsymbol\sigma(v))\cdot\boldsymbol n_K w_K\,\mathrm ds\\ &= \sum_{K\in\mathcal T_h}\int_K\big(\mathcal K_\delta^{\mathrm d}(\boldsymbol\sigma(v))\cdot\nabla w_K+\mathcal K_\delta^{\mathrm b}(\nabla\cdot\boldsymbol\sigma(v))w_K\big)\,\mathrm dx, \end{aligned}$$

where we used the divergence formula and the commuting property (40.19) in the last line. Letting $\delta\to0$, we conclude that $n_{\sharp\delta}(v,\hat w_h)$ also tends to the right-hand side of (41.28b) as $\delta\to0$. Hence, (41.28b) holds true.

(iii) Let us set $\hat\zeta_h^k:=\hat{\mathcal I}_h^k(u)-\hat u_h\in\hat V_{h,0}^k$. The coercivity of $\hat a_h$ on $\hat V_{h,0}^k$ and the definition of the consistency error imply that

$$\alpha\,\|\hat\zeta_h^k\|_{\hat V_{h,0}^k}^2\le\hat a_h(\hat\zeta_h^k,\hat\zeta_h^k)=-\langle\delta_h(\hat{\mathcal I}_h^k(u)),\hat\zeta_h^k\rangle_{(\hat V_{h,0}^k)',\hat V_{h,0}^k},$$

so that $\hat a_h(\hat\zeta_h^k,\hat\zeta_h^k)\le\alpha^{-1}\|\delta_h(\hat{\mathcal I}_h^k(u))\|_{(\hat V_{h,0}^k)'}^2$. The consistency/boundedness property from Lemma 41.16 yields

$$\hat a_h(\hat\zeta_h^k,\hat\zeta_h^k)\le c\,\|u-\mathcal E_h^{k+1}(u)\|_{V_\sharp}^2.$$

Recalling that $u_K:=u_{\mathcal T_h|K}$, $u_{\partial K}:=u_{\mathcal F_h|\partial K}$, the definitions of $\hat a_h$ and of $\hat\zeta_h^k$ imply that

$$\sum_{K\in\mathcal T_h}\lambda_K\|\nabla\mathsf R(\hat{\mathcal I}_K^k(u)-\hat u_K)\|_{\boldsymbol L^2(K)}^2\le c\,\|u-\mathcal E_h^{k+1}(u)\|_{V_\sharp}^2.$$

Since $\mathsf R(\hat{\mathcal I}_K^k(u))=\mathcal E_K(u)$ for all $K\in\mathcal T_h$, we have

$$u-\mathsf R(\hat u_K)=(u-\mathcal E_K(u))+\mathsf R(\hat{\mathcal I}_K^k(u)-\hat u_K).$$

Thus, the estimate (41.31) follows from the triangle inequality and the fact that

$$\sum_{K\in\mathcal{T}_h} \lambda_K \|\nabla(u-\mathcal{E}_K(u))\|_{\boldsymbol{L}^2(K)}^2 \le \|u-\mathcal{E}_h^{k+1}(u)\|_{V_\sharp}^2.$$

(iv) Let us set $\eta^{k+1} := u-\mathcal{E}_h^{k+1}(u)$. We need to bound $\|\eta^{k+1}\|_{V_\sharp} = |\eta^{k+1}|_{\lambda,p,q}$. Recalling (41.11), we need to estimate the terms $\|\nabla(\eta_{|K}^{k+1})\|_{\boldsymbol{L}^2(K)}$, $h_K^{d(\frac12-\frac1p)}\|\nabla(\eta_{|K}^{k+1})\|_{\boldsymbol{L}^p(K)}$, and $h_K^{d(\frac{d+2}{2d}-\frac1q)}\|\Delta(\eta_{|K}^{k+1})\|_{L^q(K)}$. We have seen in Exercise 39.3 that

$$\|\nabla(\eta_{|K}^{k+1})\|_{\boldsymbol{L}^2(K)} \le c\, h_K^t |u|_{H^{1+t}(K)}, \qquad t := \min(r,k+1).$$

Let us now consider the other two terms. Let $\ell := \lceil t\rceil$, so that $\ell\le k+1$ and $\ell\le 1+t$. Let us set $\eta^\ell := u-\mathcal{E}_h^\ell(u)$. Note that we also have $\|\nabla(\eta_{|K}^\ell)\|_{\boldsymbol{L}^2(K)} \le c h_K^t |u|_{H^{1+t}(K)}$. Invoking the triangle inequality, an inverse inequality, and the triangle inequality again, we infer that

$$h_K^{d(\frac12-\frac1p)}\|\nabla(\eta_{|K}^{k+1})\|_{\boldsymbol{L}^p(K)} \le h_K^{d(\frac12-\frac1p)}\|\nabla(\eta_{|K}^{\ell})\|_{\boldsymbol{L}^p(K)} + c\left(\|\nabla(\eta_{|K}^{k+1})\|_{\boldsymbol{L}^2(K)} + \|\nabla(\eta_{|K}^{\ell})\|_{\boldsymbol{L}^2(K)}\right),$$

and the two terms between the parentheses are bounded by $ch_K^t|u|_{H^{1+t}(K)}$. Moreover, invoking (41.16), we obtain

$$\begin{aligned} h_K^{d(\frac12-\frac1p)}\|\nabla(\eta_{|K}^{\ell})\|_{\boldsymbol{L}^p(K)} &\le c\left(\|\nabla(\eta_{|K}^{\ell})\|_{\boldsymbol{L}^2(K)} + h_K^t|\nabla(\eta_{|K}^{\ell})|_{\boldsymbol{H}^t(K)}\right)\\ &= c\left(\|\nabla(\eta_{|K}^{\ell})\|_{\boldsymbol{L}^2(K)} + h_K^t|u|_{H^{1+t}(K)}\right)\\ &\le c' h_K^t|u|_{H^{1+t}(K)}, \end{aligned}$$

since $t\le\ell$. Similarly, we have

$$h_K^{d(\frac{d+2}{2d}-\frac1q)}\|\Delta(\eta_{|K}^{k+1})\|_{L^q(K)} \le h_K^{d(\frac{d+2}{2d}-\frac1q)}\|\Delta(\eta_{|K}^{\ell})\|_{L^q(K)} + c\left(\|\nabla(\eta_{|K}^{k+1})\|_{\boldsymbol{L}^2(K)} + \|\nabla(\eta_{|K}^{\ell})\|_{\boldsymbol{L}^2(K)}\right).$$

It remains to estimate $h_K^{d(\frac{d+2}{2d}-\frac1q)}\|\Delta(\eta_{|K}^{\ell})\|_{L^q(K)}$. We proceed as in the end of the proof of Theorem 41.8. If $t\le 1$ (so that $\chi_t=1$), we have $\ell=1$, and we infer that

$$h_K^{d(\frac{d+2}{2d}-\frac1q)}\|\Delta(\eta_{|K}^{\ell})\|_{L^q(K)} = \lambda_K^{-1} h_K^{d(\frac{d+2}{2d}-\frac1q)}\|f\|_{L^q(K)}.$$

Otherwise, we have $t>1$ (so that $\chi_t=0$) and $\ell\ge 2$, and we take $q=2$. Then, using the triangle inequality, an inverse inequality, and the triangle inequality again, we obtain

$$h_K\|\Delta(\eta_{|K}^{\ell})\|_{L^q(K)} \le h_K\|\Delta(u-\Pi_K^\ell(u)\|_{L^q(K)} + c\left(\|\nabla(u-\Pi_K^\ell(u))\|_{\boldsymbol{L}^2(K)} + \|\nabla(\eta_{|K}^{\ell})\|_{\boldsymbol{L}^2(K)}\right).$$

We conclude by invoking the approximation properties of Π_K^ℓ (the L^2-orthogonal projection onto $\mathbb{P}_{\ell,d}$) and since $\|\nabla(\eta_{|K}^{\ell})\|_{\boldsymbol{L}^2(K)} \le ch_K^t|u|_{H^{1+t}(K)}$.

Chapter 42

Linear elasticity

Exercises

Exercise 42.1 (Compliance). (i) Let $\mathbb{s}(\mathbb{e})$ be defined in (42.3) (i.e., $\mathbb{s}(\mathbb{e}) := 2\mu\mathbb{e} + \lambda\mathrm{tr}(\mathbb{e})\mathbb{I}_d$) and let $\mathbb{A}$ be the fourth order tensor s.t. $\mathbb{s}(\mathbb{e}) = \mathbb{A}\mathbb{e}$. Verify that $\mathbb{A}$ is symmetric positive definite. (*Hint*: compute the quadratic form $\mathbb{A}\mathbb{e}:\mathbb{f}$.) Compute $\mathbb{A}^{\frac{1}{2}}\mathbb{e}$. (*Hint*: find $\alpha, \beta \in \mathbb{R}$ s.t. $\mathbb{A}^{\frac{1}{2}}\mathbb{e} = \alpha\mathbb{e} + \beta\,\mathrm{tr}(\mathbb{e})\mathbb{I}$.) (ii) Invert (42.3), i.e., express $\mathbb{e}$ as a function of $\mathbb{s}$ (the fourth-order tensor $\mathbb{C}$ s.t. $\mathbb{e} = \mathbb{C}\mathbb{s}$ is called *compliance tensor*). (*Hint*: compute first $\mathrm{tr}(\mathbb{s})$.) Compute $\mathbb{e}:\mathbb{s}$ in terms of $\mathbb{s}'$ and $\mathrm{tr}(\mathbb{s})$ where $\mathbb{t}' := \mathbb{t} - \frac{1}{3}\mathrm{tr}(\mathbb{t})\mathbb{I}$ is the deviatoric (i.e., trace-free) part of the tensor $\mathbb{t}$. (iii) Consider the Hellinger–Reissner functional $\mathfrak{L}_{\mathrm{HR}}(\mathbb{t}, \boldsymbol{v}) := \int_D (\frac{1}{4\mu}\mathbb{t}':\mathbb{t}' + \frac{1}{18\kappa}\,\mathrm{tr}(\mathbb{t})^2 + (\nabla\cdot\mathbb{t})\cdot\boldsymbol{v} - \boldsymbol{f}\cdot\boldsymbol{v})\,\mathrm{d}x$ on $\mathbb{H} \times \boldsymbol{V}$ where $\mathbb{H} := \{\mathbb{t} \in \mathbb{L}^2(D) \mid \mathbb{t} = \mathbb{t}^\mathsf{T}, \nabla\cdot\mathbb{t} \in \boldsymbol{L}^2(D)\}$ and $\boldsymbol{V} := \boldsymbol{L}^2(D)$. Find the equations (in weak form) satisfied by a critical point $(\mathbb{s}, \boldsymbol{u})$ of $\mathfrak{L}_{\mathrm{HR}}$. Verify that $(\mathbb{s}, \boldsymbol{u})$ satisfies (42.1) and (42.3) a.e. in D. (*Hint*: use a density argument.)

Exercise 42.2 (Second-order system). (i) Find matrices $\mathbb{A}^{jk} \in \mathbb{R}^{d\times d}$ for all $j, k \in \{1{:}d\}$ s.t. $\nabla\cdot\mathbb{s}(\boldsymbol{u}) = \sum_{j,k} \partial_j(\mathbb{A}^{jk}\partial_k\boldsymbol{u})$. (*Hint*: verify that $\sum_{j,k} \partial_j(\lambda(\boldsymbol{e}_j \otimes \boldsymbol{e}_k)\partial_k\boldsymbol{u}) = \nabla(\lambda\nabla\cdot\boldsymbol{u})$ and $\sum_{j,k} \partial_j(\mu(\boldsymbol{e}_k \otimes \boldsymbol{e}_j)\partial_k\boldsymbol{u}) = \nabla\cdot(\mu\nabla\boldsymbol{u}^\mathsf{T})$ where $(\boldsymbol{e}_j)_{j\in\{1:d\}}$ is the canonical basis of $\mathbb{R}^d$.) (ii) Verify that $(\mathbb{A}^{jk})^\mathsf{T} = \mathbb{A}^{kj}$. What is the consequence on the bilinear form $a(\boldsymbol{v}, \boldsymbol{w}) := \int_D \partial_j\boldsymbol{w}^\mathsf{T}\mathbb{A}^{jk}\partial_k\boldsymbol{v}\,\mathrm{d}x$?

Exercise 42.3 (Pure traction). The pure traction problem is $\nabla\cdot\mathbb{s}(\boldsymbol{u}) + \boldsymbol{f} = \boldsymbol{0}$ in D and $\mathbb{s}(\boldsymbol{u})\cdot\boldsymbol{n} = \boldsymbol{g}$ on ∂D. (i) Write a weak formulation in $\boldsymbol{H}^1(D)$. (ii) Show that it is necessary that $\int_D \boldsymbol{f}\cdot\boldsymbol{r}\,\mathrm{d}x + \int_{\partial D} \boldsymbol{g}\cdot\boldsymbol{r}\,\mathrm{d}s = 0$ for a weak solution to exist. (iii) Assume that $\boldsymbol{r} \in \boldsymbol{R}$ satisfies $\int_D \boldsymbol{r}\,\mathrm{d}x = \boldsymbol{0}$ and $\int_D \nabla\times\boldsymbol{r}\,\mathrm{d}x = \boldsymbol{0}$. Show that $\boldsymbol{r} = \boldsymbol{0}$. (iv) Let $\boldsymbol{V} := \{\boldsymbol{v} \in \boldsymbol{H}^1(D) \mid \int_D \boldsymbol{v}\,\mathrm{d}x = \boldsymbol{0},\ \int_D \nabla\times\boldsymbol{v}\,\mathrm{d}x = \boldsymbol{0}\}$. Show that the weak formulation is well-posed in $\boldsymbol{V}$.

Exercise 42.4 (Timoshenko beam). Consider a horizontal beam $D := (0, L)$ clamped at $x = 0$ and subjected to a (vertical) force distribution f and to a bending moment distribution m. A (vertical) shear force F and a bending moment M are applied at $x = L$. The unknowns are the vertical displacement u and the rotation angle of the transverse section θ s.t. $-(u'' - \theta') = \frac{\gamma}{EI}f$ and $-\gamma\theta'' - (u' - \theta) = \frac{\gamma}{EI}m$ in D, where E is the Young modulus, I is the area moment of inertia, and $\gamma := \frac{2(1+\nu)I}{S\kappa}$ (S is the cross section area and κ is an empirical correction factor usually set to $\frac{5}{6}$). The boundary conditions are $u(0) = 0$, $\theta(0) = 0$, $(u' - \theta)(L) = \frac{\gamma}{EI}F$, and $\theta'(L) = \frac{1}{EI}M$. (i) Assuming $f, m \in L^2(D)$, write a weak formulation for the pair (u, θ) in $Y := X \times X$ with $X := \{v \in H^1(D) \mid v(0) = 0\}$. (ii) Prove the well-posedness of the weak formulation. (*Hint*: use

that $2\int_D \theta u' \,\mathrm{d}x \le \mu\|\theta\|_{L^2(D)}^2 + \frac{1}{\mu}|u|_{H^1(D)}^2$ with μ sufficiently close to 1 and the Poincaré–Steklov inequality.) (iii) Write an H^1-conforming finite element approximation and derive H^1- and L^2-error estimates for u and θ.

Exercise 42.5 (HHO). (i) Prove (42.25). (ii) Prove Lemma 42.20. (*Hint*: see Lemma 39.2 and use the local Korn inequality $\|\boldsymbol{v}\|_{\boldsymbol{L}^2(K)} \le ch_K\|\mathfrak{e}(\boldsymbol{v})\|_{\mathbb{L}^2(K)}$ for all $\boldsymbol{v} \in \boldsymbol{H}^1(K)$ s.t. $(\boldsymbol{v},\boldsymbol{r})_{\boldsymbol{L}^2(K)} = 0$ for all $\boldsymbol{r} \in \boldsymbol{R}_K$; see Horgan [28], Kim [32].) (iii) Prove Lemma 42.21. (*Hint*: adapt the proof of Lemma 39.16.)

Solution to exercises

Exercise 42.1 (Compliance). (i) We have

$$\mathbb{A}\mathfrak{e}:\mathfrak{f} = (2\mu\mathfrak{e} + \lambda\,\mathrm{tr}(\mathfrak{e})\mathbb{I}):\mathfrak{f} = 2\mu\mathfrak{e}:\mathfrak{f} + \lambda\,\mathrm{tr}(\mathfrak{e})\,\mathrm{tr}(\mathfrak{f}).$$

This expression is symmetric in $\mathfrak{e}$ and $\mathfrak{f}$. Reasoning as in the proof of Theorem 42.11, we also have

$$\mathbb{A}\mathfrak{e}:\mathfrak{e} = 2\mu\mathfrak{e}:\mathfrak{e} + \lambda\,\mathrm{tr}(\mathfrak{e})^2 \ge \min(2\mu, 3\kappa)\mathfrak{e}:\mathfrak{e},$$

where μ and κ are bounded from below away from zero. Hence, $\mathbb{A}\mathfrak{e}:\mathfrak{e} \ge 0$ and $\mathbb{A}\mathfrak{e}:\mathfrak{e} = 0$ implies that $\mathfrak{e}$ is zero. Using the ansatz $\mathbb{A}^{\frac12}\mathfrak{e} = \alpha\mathfrak{e} + \beta\,\mathrm{tr}(\mathfrak{e})\mathbb{I}$ and recalling that $d = 3$, we have

$$\mathbb{A}\mathfrak{e}:\mathfrak{e} = \mathbb{A}^{\frac12}\mathfrak{e}:\mathbb{A}^{\frac12}\mathfrak{e} = \alpha^2\mathfrak{e}:\mathfrak{e} + (2\alpha\beta + 3\beta^2)\,\mathrm{tr}(\mathfrak{e})^2.$$

We identify the coefficients with the above expression of $\mathbb{A}\mathfrak{e}:\mathfrak{e}$ and infer that

$$\alpha^2 = 2\mu, \qquad 3\beta^2 + 2\alpha\beta = \lambda.$$

Hence, $\alpha = \sqrt{2\mu}$ and $\beta = \frac13(\sqrt{3\kappa} - \sqrt{2\mu})$.
(ii) Taking the trace of (42.3), we infer that

$$\mathrm{tr}(\mathbb{s}) = (2\mu + 3\lambda)\,\mathrm{tr}(\mathfrak{e}) = 3\kappa\,\mathrm{tr}(\mathfrak{e}).$$

Since $\kappa > 0$ by assumption, we have $\mathrm{tr}(\mathfrak{e}) = \frac{1}{3\kappa}\,\mathrm{tr}(\mathbb{s})$, so that $\mathbb{s} = 2\mu\mathfrak{e} + \frac{\lambda}{3\kappa}\,\mathrm{tr}(\mathbb{s})\mathbb{I}$. Since $\mu > 0$ by assumption, we conclude that

$$\mathfrak{e} = \frac{1}{2\mu}\left(\mathbb{s} - \frac{\lambda}{3\kappa}\,\mathrm{tr}(\mathbb{s})\mathbb{I}\right).$$

Introducing the deviatoric part of $\mathbb{s}$, i.e., $\mathbb{s}' := \mathbb{s} - \frac13\,\mathrm{tr}(\mathbb{s})\mathbb{I}$, we have

$$2\mu\mathfrak{e} = \mathbb{s}' + \frac{2\mu}{9\kappa}\,\mathrm{tr}(\mathbb{s})\mathbb{I}.$$

Since $\mathbb{s}':\mathbb{I} = \mathrm{tr}(\mathbb{s}') = 0$ and $\mathbb{I}:\mathbb{I} = 3$, we obtain

$$\mathfrak{e}:\mathbb{s} = \frac{1}{2\mu}\mathbb{s}':\mathbb{s}' + \frac{1}{9\kappa}\,\mathrm{tr}(\mathbb{s})^2.$$

(iii) The Fréchet derivative of the functional $\mathfrak{L}_{\mathrm{HR}}$ at a critical point $(\mathbb{s},\boldsymbol{u})$ is s.t.

$$\partial_{\mathbb{t}}\mathfrak{L}_{\mathrm{HR}}(\mathbb{s},\boldsymbol{u})(\mathbb{h}) = \int_D \left(\tfrac{1}{2\mu}\mathbb{s}':\mathbb{h}' + \tfrac{1}{9\kappa}\,\mathrm{tr}(\mathbb{s})\,\mathrm{tr}(\mathbb{h}) + (\nabla\cdot\mathbb{h})\cdot\boldsymbol{u}\right)\mathrm{d}x,$$

$$\partial_{\boldsymbol{v}}\mathfrak{L}_{\mathrm{HR}}(\mathbb{s},\boldsymbol{u})(\boldsymbol{g}) = \int_D (\nabla\cdot\mathbb{s} - \boldsymbol{f})\cdot\boldsymbol{g}\,\mathrm{d}x,$$

for all $(\mathbb{h}, \boldsymbol{g}) \in \mathbb{H} \times \boldsymbol{V}$. In the first equation, we observe that

$$\begin{aligned}\frac{1}{2\mu}\mathbb{s}':\mathbb{h}' + \frac{1}{9\kappa}\operatorname{tr}(\mathbb{s})\operatorname{tr}(\mathbb{h}) &= \frac{1}{2\mu}\mathbb{s}':\mathbb{h} + \frac{1}{9\kappa}\operatorname{tr}(\mathbb{s})\operatorname{tr}(\mathbb{h}) \\ &= \frac{1}{2\mu}(\mathbb{s} - \frac{1}{3}\operatorname{tr}(\mathbb{s})\mathbb{I}):\mathbb{h} + \frac{1}{9\kappa}\operatorname{tr}(\mathbb{s})\mathbb{I}:\mathbb{h} \\ &= \frac{1}{2\mu}\left(\mathbb{s} - \frac{\lambda}{3\kappa}\operatorname{tr}(\mathbb{s})\mathbb{I}\right):\mathbb{h}.\end{aligned}$$

Taking $\mathbb{h}$ to be smooth and compactly supported in D, and recalling that $\mathbb{h}$ takes symmetric values so that $\int_D (\nabla\cdot\mathbb{h})\cdot\boldsymbol{u} = -\int_D \mathbb{e}(u):\mathbb{h}\,\mathrm{d}x$ (in the weak sense), we infer that

$$\frac{1}{2\mu}\left(\mathbb{s} - \frac{\lambda}{3\kappa}\operatorname{tr}(\mathbb{s})\mathbb{I}\right) = \mathbb{e}(\boldsymbol{u}).$$

Step (ii) shows that if $(\mathbb{s}, \boldsymbol{u})$ is a critical point of $\mathfrak{L}_{\mathrm{HR}}$, then $(\mathbb{s}, \boldsymbol{u})$ satisfies the constitutive relation (42.3) a.e. in D. Finally, that the equilibrium condition (42.1) is satisfied a.e. in D follows by taking $\boldsymbol{g}$ to be smooth and compactly supported in D.

Exercise 42.2 (Second-order system). (i) To verify the hint, we observe that

$$\sum_{j,k}\partial_j(\lambda(\boldsymbol{e}_j\otimes\boldsymbol{e}_k)\partial_k\boldsymbol{u}) = \sum_{j,k}\partial_j(\lambda\boldsymbol{e}_j\partial_k u_k) = \sum_j \partial_j(\lambda\boldsymbol{e}_j\nabla\cdot\boldsymbol{u}) = \nabla(\lambda\nabla\cdot\boldsymbol{u}),$$

and that

$$\sum_{j,k}\partial_j(\mu(\boldsymbol{e}_k\otimes\boldsymbol{e}_j)\partial_k\boldsymbol{u}) = \sum_{j,k}\partial_j(\mu\boldsymbol{e}_k\partial_k u_j) = \nabla\cdot(\mu\nabla\boldsymbol{u}^{\mathsf{T}}).$$

Furthermore, we have $\sum_{j,k}\partial_j(\mu\delta_{jk}\mathbb{I}\partial_k\boldsymbol{u}) = \nabla\cdot(\mu\nabla\boldsymbol{u})$. We conclude that

$$\mathbb{A}^{jk} = \mu\delta_{jk}\mathbb{I} + \mu\boldsymbol{e}_k\otimes\boldsymbol{e}_j + \lambda\boldsymbol{e}_j\otimes\boldsymbol{e}_k.$$

(ii) It is clear that $(\mathbb{A}^{jk})^{\mathsf{T}} = \mathbb{A}^{kj}$. This implies that the bilinear form $a(\boldsymbol{v},\boldsymbol{w}) := \int_D \partial_j\boldsymbol{w}^{\mathsf{T}}\mathbb{A}^{jk}\partial_k\boldsymbol{v}\,\mathrm{d}x$ is symmetric.

Exercise 42.3 (Pure traction). (i) Le $\boldsymbol{v}$ be a smooth test function. By proceeding as in §42.2.1, we obtain

$$\begin{cases}\text{Find } \boldsymbol{u}\in\boldsymbol{H}^1(D) \text{ such that}\\ a(\boldsymbol{u},\boldsymbol{w}) = \int_D \boldsymbol{f}\cdot\boldsymbol{w}\,\mathrm{d}x + \int_{\partial D}\boldsymbol{g}\cdot\boldsymbol{w}\,\mathrm{d}s, \quad \forall\boldsymbol{w}\in\boldsymbol{H}^1(D),\end{cases}$$

where $a(\boldsymbol{v},\boldsymbol{w}) = \int_D \big(2\mu\mathbb{e}(\boldsymbol{v}):\mathbb{e}(\boldsymbol{w}) + \lambda(\nabla\cdot\boldsymbol{v})(\nabla\cdot\boldsymbol{w})\big)\,\mathrm{d}x$.
(ii) Observe first $\mathbb{e}(\boldsymbol{r}) = 0$ and $\nabla\cdot\boldsymbol{r} = 0$ for all $\boldsymbol{r}\in\boldsymbol{R}$. Assume that the above problem has a solution $\boldsymbol{u}\in\boldsymbol{H}^1(D)$. Using test functions in $\boldsymbol{R}\subset\boldsymbol{H}^1(D)$, we infer that $0 = a(\boldsymbol{u},\boldsymbol{r}) = \int_D \boldsymbol{f}\cdot\boldsymbol{r}\,\mathrm{d}x + \int_{\partial D}\boldsymbol{g}\cdot\boldsymbol{r}\,\mathrm{d}s$ for all $\boldsymbol{r}\in\boldsymbol{R}$. Hence, it is necessary that the volume and surface loads $\boldsymbol{f}$ and $\boldsymbol{g}$ satisfy the above compatibility equation for a weak solution to exist in $\boldsymbol{H}^1(D)$.
(iii) Let $\boldsymbol{r} = \boldsymbol{\alpha} + \boldsymbol{\beta}\times\boldsymbol{x}\in\boldsymbol{R}$ and assume that $\int_D \boldsymbol{r}\,\mathrm{d}x = \boldsymbol{0}$ and $\int_D \nabla\times\boldsymbol{r}\,\mathrm{d}x = \boldsymbol{0}$. Observing that $\nabla\times\boldsymbol{r} = 2\boldsymbol{\beta}$, the condition $\int_D\nabla\times\boldsymbol{r}\,\mathrm{d}x = \boldsymbol{0}$ implies that $\boldsymbol{\beta} = \boldsymbol{0}$. The condition $\int_D\boldsymbol{r}\,\mathrm{d}x = \boldsymbol{\alpha}\int_D\mathrm{d}x = \boldsymbol{0}$ implies that $\boldsymbol{\alpha} = \boldsymbol{0}$. In conclusion, $\boldsymbol{r} = \boldsymbol{0}$.
(iv) Let $\boldsymbol{V} := \{\boldsymbol{v}\in\boldsymbol{H}^1(D) \mid \int_D\boldsymbol{v}\,\mathrm{d}x = \boldsymbol{0},\ \int_D\nabla\times\boldsymbol{v}\,\mathrm{d}x = \boldsymbol{0}\}$ and consider the weak problem: Find $\boldsymbol{u}\in\boldsymbol{V}$ such that $a(\boldsymbol{u},\boldsymbol{w}) = \int_D\boldsymbol{f}\cdot\boldsymbol{w}\,\mathrm{d}x + \int_{\partial D}\boldsymbol{g}\cdot\boldsymbol{w}\,\mathrm{d}s$ for all $\boldsymbol{w}\in\boldsymbol{V}$. The well-posedness of this problem is established by proceeding as in the proof of Theorem 42.11. The boundedness of a and of the right-hand side is evident. To prove the coercivity of the bilinear form a, we use that $\boldsymbol{V}\cap\boldsymbol{R} = \{\boldsymbol{0}\}$ and apply Korn's second inequality (see (42.14)).

Exercise 42.4 (Timoshenko beam). (i) Let v be a test function for the normal displacement, and let ω be a test function for the rotation angle. Multiplying the first equation by v, the second by ω, and integrating by parts over D, we obtain $a((u,\theta),(v,\omega)) = \ell(v,\omega)$ with

$$a((u,\theta),(v,\omega)) := \int_D \gamma\theta'\omega' \,\mathrm{d}x + \int_D (u'-\theta)(v'-\omega)\,\mathrm{d}x,$$
$$\ell(v,\omega) := \frac{\gamma}{EI}\left(\int_D (fv+m\omega)\,\mathrm{d}x + Fv(L) + M\omega(L)\right).$$

To make sense of the above integrals, we introduce the Hilbert space $X := \{v \in H^1(D) \mid v(0) = 0\}$, and we equip the product space $X\times X$ with the norm $\|(u,\theta)\|_{X\times X} := |u|_{H^1(D)} + L|\theta|_{H^1(D)}$. Then, one possible weak formulation of the problem is as follows: Find $(u,\theta) \in X\times X$ such that $a((u,\theta),(v,\omega)) = \ell(v,\omega)$ for all $(v,\omega) \in X\times X$.
(ii) The boundedness of a and ℓ is an application of the Cauchy–Schwarz inequality. Let us prove the coercivity of a. A straightforward calculation yields

$$a\big((u,\theta),(u,\theta)\big) = \int_D \gamma|\theta'|^2\,\mathrm{d}x + \int_D |u'|^2\,\mathrm{d}x + \int_D \theta^2\,\mathrm{d}x - 2\int_D \theta u'\,\mathrm{d}x.$$

Let $\mu > 0$. Using the arithmetic-geometric inequality (C.5) with parameter μ, together with the Poincaré–Steklov inequality $C_{\mathrm{PS}}\|v\|_{L^2(D)} \le L\|v'\|_{L^2(D)}$ valid for all $v \in X$, we obtain (with the nondimensional number $\tilde\gamma = L^{-2}\gamma$)

$$\begin{aligned} a\big((u,\theta),(u,\theta)\big) &\ge \gamma|\theta|^2_{H^1(D)} + |u|^2_{H^1(D)} + \|\theta\|^2_{L^2(D)} - \mu\|\theta\|^2_{L^2(D)} - \frac{1}{\mu}|u|^2_{H^1(D)} \\ &\ge \left(1-\frac{1}{\mu}\right)|u|^2_{H^1(D)} + \frac{\gamma}{2}|\theta|^2_{H^1(D)} + \left(\frac{\tilde\gamma}{2}C_{\mathrm{PS}}^2 + 1 - \mu\right)\|\theta\|^2_{L^2(D)}. \end{aligned}$$

Taking $\mu := 1 + \frac{\tilde\gamma}{2}C_{\mathrm{PS}}^2$ yields

$$a\big((u,\theta),(u,\theta)\big) \ge \frac{\frac{\tilde\gamma}{2}C_{\mathrm{PS}}^2}{1+\frac{\tilde\gamma}{2}C_{\mathrm{PS}}^2}|u|^2_{H^1(D)} + \frac{\tilde\gamma}{2}L^2|\theta|^2_{H^1(D)} \ge \alpha(\tilde\gamma)\|(u,\theta)\|^2_{X\times X},$$

with $\alpha(\tilde\gamma) := \frac{\tilde\gamma}{2}\inf\big(1, C_{\mathrm{PS}}^2/(1+\frac{\tilde\gamma}{2}C_{\mathrm{PS}}^2)\big) > 0$. This proves that a is coercive since $\gamma > 0$. Owing to the Lax–Milgram lemma, we infer the well-posedness of the weak formulation. Since the weak solution $(u,\theta) \in X\times X$ satisfies both PDEs in $L^2(D)$, we have $u'' = \theta' - \frac{\gamma}{EI}f$ and $\theta'' = -\frac{1}{\gamma}(u'-\theta) - \frac{1}{EI}m$, which immediately shows that both u and θ are in $H^2(D)$. Finally, since a is symmetric, Proposition 25.8 shows that $(u,\theta) = \arg\inf_{(v,\omega)\in Y}\mathfrak{E}(v,\omega)$ with

$$\mathfrak{E}(v,\omega) := \int_D \frac{1}{2}(\gamma|\omega'|^2 + |v'-\omega|^2)\,\mathrm{d}x - \frac{\gamma}{EI}\left(\int_D (fv+m\omega)\,\mathrm{d}x + Fv(L) + M\omega(L)\right).$$

(iii) Let $\mathcal{T}_h$ be a mesh of D with vertices $0 =: x_0 < x_1 < \ldots < x_I < x_{I+1} := L$ with $I \in \mathbb{N}$. We construct an H^1-conforming approximation space by using $\mathbb{P}_k$ Lagrange finite elements for both u and θ and by setting

$$X_h := \{v_h \in C^0(\overline{D}) \mid v_{h|[x_i,x_{i+1}]} \in \mathbb{P}_k, \forall i \in \{0{:}I\},\ v_h(0) = 0\}.$$

The discrete problem consists of seeking $(u_h,\theta_h) \in X_h\times X_h$ such that $a((u_h,\theta_h),(v_h,\omega_h)) = \ell(v_h,\omega_h)$ for all $(v_h,\omega_h) \in X_h\times X_h$. Assuming that $u,\theta \in H^{1+r}(D)$, $r \in \{1{:}k\}$, and using Céa's lemma, we infer that

$$|u-u_h|_{H^1(D)} + L|\theta-\theta_h|_{H^1(D)} \le c\,h^r(|u|_{H^{1+r}(D)} + L|\theta|_{H^{1+r}(D)}).$$

We can also apply the Aubin–Nitsche Lemma since the elliptic regularity theory leads to the pickup index $s = 1$. This yields

$$\|u - u_h\|_{L^2(D)} + L\|\theta - \theta_h\|_{L^2(D)} \le c\, h^{1+r}(|u|_{H^{1+r}(D)} + L|\theta|_{H^{1+r}(D)}).$$

Exercise 42.5 (HHO). (i) For all $q \in V_K^k$, integrating by parts in (42.24), we have

$$\begin{aligned}(\mathsf{D}(\hat{\mathcal{I}}_K^k(\boldsymbol{v})), q)_{L^2(K)} &= -(\Pi_K^k(\boldsymbol{v}), \nabla q)_{L^2(K)} + (\Pi_{\partial K}^k(\boldsymbol{v}_{\partial K}), q\boldsymbol{n}_K)_{\boldsymbol{L}^2(K)} \\ &= -(\boldsymbol{v}, \nabla q)_{L^2(K)} + (\boldsymbol{v}_{\partial K}, q\boldsymbol{n}_K)_{\boldsymbol{L}^2(K)} = (\nabla{\cdot}\boldsymbol{v}, q)_{L^2(K)},\end{aligned}$$

since $\nabla q \in \boldsymbol{V}_K^k$ and $q\boldsymbol{n}_K \in \boldsymbol{V}_{\partial K}^k$. Since $\mathsf{D}(\hat{\mathcal{I}}_K^k(\boldsymbol{v})) \in V_K^k$ by definition of D, we conclude that (42.25) holds true.
(ii) The only difference with the proof of Lemma 39.2 is that instead of the local Poincaré–Steklov inequality in K, we invoke the local Korn inequality mentioned in the hint, i.e.,, $\|\boldsymbol{v}\|_{\boldsymbol{L}^2(K)} \le c h_K \|\mathfrak{e}(\boldsymbol{v})\|_{\mathbb{L}^2(K)}$ for all $\boldsymbol{v} \in \boldsymbol{H}^1(K)$ s.t. $(\boldsymbol{v}, \boldsymbol{r})_{\boldsymbol{L}^2(K)} = 0$ for all $\boldsymbol{r} \in \boldsymbol{R}_K$. The assumption $k \ge 1$ is used here to prove that $((I - \Pi_K^k)\mathsf{R}(\hat{\boldsymbol{v}}_K), \boldsymbol{r})_{\boldsymbol{L}^2(K)} = 0$ for all $\boldsymbol{r} \in \boldsymbol{R}_K \subset \mathbb{P}_{1,d} \circ \boldsymbol{T}_K^{-1}$.
(iii) The proof is similar to that of Lemma 39.16 for scalar elliptic PDEs. We obtain

$$\langle \delta_{\mathcal{I}}(\boldsymbol{u}), \hat{\boldsymbol{w}}_h \rangle_{(\hat{\boldsymbol{V}}_{h,0}^k)', \hat{\boldsymbol{V}}_{h,0}^k} = -\sum_{K \in \mathcal{T}_h} \big(\mathfrak{T}_{1,K} + \mathfrak{T}_{2,K} + \mathfrak{T}_{3,K}\big),$$

where

$$\begin{aligned}\mathfrak{T}_{1,K} &:= \mu(\mathfrak{e}(\boldsymbol{u} - \boldsymbol{\mathcal{E}}_K(\boldsymbol{u}))\boldsymbol{n}_K, \boldsymbol{w}_K - \boldsymbol{w}_{\partial K})_{\boldsymbol{L}^2(\partial K)}, \\ \mathfrak{T}_{2,K} &:= \lambda(\nabla{\cdot}\boldsymbol{u} - \mathsf{D}(\hat{\mathcal{I}}_K^k(\boldsymbol{u})), \boldsymbol{w}_K - \boldsymbol{w}_{\partial K})_{L^2(\partial K)}, \\ \mathfrak{T}_{3,K} &:= \mu h_K^{-1}(\mathsf{S}(\hat{\mathcal{I}}_K^k(\boldsymbol{u})), \mathsf{S}(\hat{\boldsymbol{w}}_K))_{\boldsymbol{L}^2(\partial K)}.\end{aligned}$$

The first and third terms are similar to those obtained in Lemma 39.16 and are estimated in the same manner, the only difference being that we invoke Korn's inequality in each cell $K \in \mathcal{T}_h$ instead of the Poincaré–Steklov inequality when estimating $\mathfrak{T}_{3,K}$. The term $\mathfrak{T}_{2,K}$ is rewritten using the commuting property (42.25) and is bounded by using the Cauchy–Schwarz inequality.

Chapter 43

Maxwell's equations: $\boldsymbol{H}(\mathrm{curl})$-approximation

Exercises

Exercise 43.1 (Compactness). Let $D := (0,1)^3$ be the unit cube in $\mathbb{R}^3$. Show that the embedding $\boldsymbol{H}_0(\mathrm{curl};D) \hookrightarrow \boldsymbol{L}^2(D)$ is not compact. (*Hint*: consider $\boldsymbol{v}_n := \nabla\phi_n$ with $\phi_n(x_1,x_2,x_3) := \frac{1}{n\pi}\sin(n\pi x_1)\sin(n\pi x_2)\sin(n\pi x_3)$, $n \geq 1$, and prove first that $(\boldsymbol{v}_n)_{n\geq 1}$ weakly converges to zero in $\boldsymbol{L}^2(D)$ (see Definition C.28), then compute $\|\boldsymbol{v}_n\|_{\boldsymbol{L}^2(D)}$ and argue by contradiction.)

Exercise 43.2 (Curl). (i) Let $\boldsymbol{v}$ be a smooth field. Show that $\|\nabla\times\boldsymbol{v}\|_{\ell^2}^2 \leq 2\nabla\boldsymbol{v}{:}\nabla\boldsymbol{v}$. (*Hint*: relate $\nabla\times\boldsymbol{v}$ to the components of $(\nabla\boldsymbol{v} - \nabla\boldsymbol{v}^{\mathsf{T}})$.) (ii) Show that $\|\nabla\times\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \leq |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}$ for all $\boldsymbol{v} \in \boldsymbol{H}_0^1(D)$. (*Hint*: use an integration by parts.)

Exercise 43.3 (Property (43.12)). Prove the claim in Example 43.2, i.e., for $[\theta_{\min},\theta_{\max}] \subset (-\pi,\pi)$ with $\delta := \theta_{\max} - \theta_{\min} < \pi$, letting $\theta := -\frac{1}{2}(\theta_{\min}+\theta_{\max})\frac{\pi}{2\pi-\delta}$, prove that $\theta \in (-\frac{\pi}{2},\frac{\pi}{2})$ and $[\theta_{\min}+\theta,\theta_{\max}+\theta] \subset (-\frac{\pi}{2},\frac{\pi}{2})$.

Exercise 43.4 (Dirichlet/Neumann). Let $\boldsymbol{v}$ be a smooth vector field in D such that $\boldsymbol{v}_{|\partial D_{\mathrm{d}}}\times\boldsymbol{n} = \boldsymbol{0}$. Prove that $(\nabla\times\boldsymbol{v})_{|\partial D_{\mathrm{d}}}\cdot\boldsymbol{n} = 0$. (*Hint*: compute $\int_D(\nabla\times\boldsymbol{v})\cdot\nabla q\,\mathrm{d}x$ with q well chosen.)

Solution to exercises

Exercise 43.1 (Compactness). Let $\boldsymbol{v}_n := \nabla\phi_n$ with

$$\phi_n(x_1,x_2,x_3) := \frac{1}{n\pi}\sin(n\pi x_1)\sin(n\pi x_2)\sin(n\pi x_3), \quad n \geq 1.$$

Clearly, $\boldsymbol{v}_n \in \boldsymbol{C}^\infty(D)$ and $\boldsymbol{v}_{n|\partial D}\times\boldsymbol{n} = \boldsymbol{0}$. Hence, $\boldsymbol{v}_n \in \boldsymbol{H}_0(\mathrm{curl};D)$. Observe also that $\|\boldsymbol{v}_n\|_{\boldsymbol{L}^2(D)} = (\frac{3}{8})^{\frac{1}{2}}$ and $\nabla\times\boldsymbol{v}_n = \boldsymbol{0}$. Hence, $\|\boldsymbol{v}_n\|_{\boldsymbol{H}(\mathrm{curl};D)} = (\frac{3}{8})^{\frac{1}{2}}$, which means that the sequence $(\boldsymbol{v}_n)_{n\geq 1}$ is bounded in $\boldsymbol{H}_0(\mathrm{curl};D)$. Let us prove that the sequence $(\boldsymbol{v}_n)_{n\geq 1}$ converges weakly to zero in $\boldsymbol{L}^2(D)$. For all $\boldsymbol{\phi} \in \boldsymbol{C}_0^\infty(D)$, we have

$$(\boldsymbol{v}_n,\boldsymbol{\phi})_{\boldsymbol{L}^2(D)} = -(\phi_n,\nabla\cdot\boldsymbol{\phi})_{L^2(D)} \to 0 \quad \text{as } n\to\infty.$$

Let now $\boldsymbol{w} \in \boldsymbol{L}^2(D)$. Owing to Theorem 1.38, for all $\epsilon > 0$, there is $\boldsymbol{\phi} \in \boldsymbol{L}^2(D)$ s.t. $\|\boldsymbol{w} - \boldsymbol{\phi}\|_{\boldsymbol{L}^2(D)} \le \epsilon$. Writing $(\boldsymbol{v}_n, \boldsymbol{w})_{\boldsymbol{L}^2(D)} = (\boldsymbol{v}_n, \boldsymbol{\phi})_{\boldsymbol{L}^2(D)} + (\boldsymbol{v}_n, \boldsymbol{w} - \boldsymbol{\phi})_{\boldsymbol{L}^2(D)}$ and using the Cauchy–Schwarz inequality to bound the second term, we infer that $\limsup_{n\to\infty} |(\boldsymbol{v}_n, \boldsymbol{w})_{\boldsymbol{L}^2(D)}| \le (\frac{3}{8})^{\frac12}\epsilon$, and since $\epsilon > 0$ is arbitrary, we conclude that $\lim_{n\to\infty}(\boldsymbol{v}_n, \boldsymbol{w})_{\boldsymbol{L}^2(D)} = 0$. We have thus shown that the sequence $(\boldsymbol{v}_n)_{n\ge1}$ converges weakly to zero in $\boldsymbol{L}^2(D)$. We can now prove that the embedding $\boldsymbol{H}_0(\mathrm{curl}; D) \hookrightarrow \boldsymbol{L}^2(D)$ is not compact. Indeed, if the embedding were compact, there would exist a subsequence $(\boldsymbol{v}_{n_k})_{k\ge1}$ strongly converging to some $\boldsymbol{v} \in \boldsymbol{L}^2(D)$, but strong convergence implies weak convergence so that $\boldsymbol{v} = \boldsymbol{0}$, and $\|\boldsymbol{v}_{n_k}\|_{\boldsymbol{L}^2(D)} = (\frac{3}{8})^{\frac12}$ with strong convergence would also imply $\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} = (\frac{3}{8})^{\frac12} > 0$, which is a contradiction.

Exercise 43.2 (Curl). (i) We have

$$\|\nabla\times\boldsymbol{v}\|_{\ell^2}^2 = \frac12(\nabla\boldsymbol{v} - \nabla\boldsymbol{v}^{\mathsf{T}}):(\nabla\boldsymbol{v} - \nabla\boldsymbol{v}^{\mathsf{T}}) = \nabla\boldsymbol{v}:\nabla\boldsymbol{v} - \nabla\boldsymbol{v}:\nabla\boldsymbol{v}^{\mathsf{T}} \le 2\nabla\boldsymbol{v}:\nabla\boldsymbol{v},$$

where the last bound follows from the Cauchy–Schwarz inequality.
(ii) Let $\boldsymbol{v} \in \boldsymbol{H}_0^1(D)$. The above identity shows that $\|\nabla\times\boldsymbol{v}\|_{\boldsymbol{L}^2(D)}^2 = |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}^2 - (\nabla\boldsymbol{v}, \nabla\boldsymbol{v}^{\mathsf{T}})_{\boldsymbol{L}^2(D)}$. Using that $\boldsymbol{v}$ vanishes at the boundary, integration by parts shows that $(\nabla\boldsymbol{v}, \nabla\boldsymbol{v}^{\mathsf{T}})_{\boldsymbol{L}^2(D)} = \|\nabla\cdot\boldsymbol{v}\|_{L^2(D)}^2$. This in turn implies that $\|\nabla\times\boldsymbol{v}\|_{\boldsymbol{L}^2(D)}^2 \le |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}^2$ for all $\boldsymbol{v} \in \boldsymbol{H}_0^1(D)$.

Exercise 43.3 (Property (43.12)**).** Let us set $m := \frac12(\theta_{\min} + \theta_{\max})$. We have the following equivalences:

- $\theta < \frac{\pi}{2}$ iff $-2m < 2\pi - \delta$ iff $-\theta_{\min} < \pi$ which holds true by assumption;
- $\theta > -\frac{\pi}{2}$ iff $2m < 2\pi - \delta$ iff $\theta_{\max} < \pi$ which holds true by assumption;
- $\theta_{\max} + \theta < \frac{\pi}{2}$ iff $\frac{\delta}{2} + m\frac{\pi-\delta}{2\pi-\delta} < \frac{\pi}{2}$ iff $m < 2\pi - \delta$ iff $\theta_{\max} < \pi$ which holds true by assumption;
- $\theta_{\min} + \theta > -\frac{\pi}{2}$ iff $-\frac{\delta}{2} + m\frac{\pi-\delta}{2\pi-\delta} > -\frac{\pi}{2}$ iff $-2m < 2\pi - \delta$ iff $-\theta_{\min} < \pi$ which holds true by assumption.

Exercise 43.4 (Dirichlet/Neumann). Let $\alpha \in \widetilde{H}^{\frac12}(\partial D_{\mathrm{d}})$, which means that $\alpha \in H^{\frac12}(\partial D_{\mathrm{d}})$ and that the zero extension of α over ∂D is in $H^{\frac12}(\partial D)$. Let $q \in H^1(D)$ be the solution to the following problem

$$\Delta q = 0, \quad q_{|\partial D_{\mathrm{d}}} = \alpha, \quad q_{|\partial D_{\mathrm{n}}} = 0.$$

We infer that

$$\int_D \nabla\times\boldsymbol{v}\cdot\nabla q\,\mathrm{d}x = -\int_{\partial D}(\boldsymbol{v}\times\boldsymbol{n})\cdot\nabla q\,\mathrm{d}s = -\int_{\partial D_{\mathrm{n}}}\boldsymbol{v}\cdot(\boldsymbol{n}\times\nabla q)\,\mathrm{d}s.$$

Observe that $\boldsymbol{n}\times\nabla q_{|\partial D_{\mathrm{n}}} = \boldsymbol{0}$ since $q_{|\partial D_{\mathrm{n}}} = 0$. Hence, we have

$$0 = \int_D \nabla\times\boldsymbol{v}\cdot\nabla q\,\mathrm{d}x = \int_{\partial D}(\nabla\times\boldsymbol{v}\cdot\boldsymbol{n})q\,\mathrm{d}s = \int_{\partial D_{\mathrm{d}}}(\nabla\times\boldsymbol{v}\cdot\boldsymbol{n})\alpha\,\mathrm{d}s.$$

Since this is true for every $\alpha \in \widetilde{H}^{\frac12}(\partial D_{\mathrm{d}})$, this means that $(\nabla\times\boldsymbol{v})_{|\partial D_{\mathrm{d}}}\cdot\boldsymbol{n} = \boldsymbol{0}$.

Chapter 44

Maxwell's equations: control on the divergence

Exercises

Exercise 44.1 (Gradient). Let $\phi \in H^1_0(D)$. Prove that $\nabla\phi \in \boldsymbol{H}_0(\mathrm{curl};D)$

Exercise 44.2 (Vector potential). Let $\boldsymbol{v} \in \boldsymbol{L}^2(D)$ with $(\nu\boldsymbol{v},\nabla m_h)_{\boldsymbol{L}^2(D)} = 0$ for all $m_h \in M_{h0}$. Prove that $(\nu\boldsymbol{v},\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)} = (\nabla\times\boldsymbol{z}_h,\nabla\times\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{w}_h \in \boldsymbol{V}_{h0}$, where $\boldsymbol{z}_h$ solves a curl-curl problem on $\boldsymbol{X}_{h0\nu}$.

Exercise 44.3 (Neumann condition). Recall Remark 44.10. Assume that D is simply connected so that there is $\hat{C}_{\mathrm{PS}} > 0$ such that $\hat{C}_{\mathrm{PS}}\ell_D^{-1}\|\boldsymbol{b}\|_{\boldsymbol{L}^2(D)} \le \|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{b} \in \boldsymbol{X}_{*\nu}$. Prove that there is $\hat{C}'_{\mathrm{PS}} > 0$ such that $\hat{C}'_{\mathrm{PS}}\ell_D^{-1}|\boldsymbol{b}_h\|_{\boldsymbol{L}^2(D)} \le \|\nabla\times\boldsymbol{b}_h\|_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{b}_h \in \boldsymbol{X}_{h*\nu}$. (*Hint*: adapt the proof of Theorem 44.6 using $\mathcal{J}^{\mathrm{c}}_h$.)

Exercise 44.4 (Discrete Poincaré–Steklov for $\nabla\cdot$). Let ν be as in §44.1.1. Let $\boldsymbol{Y}_{0\nu} := \{\boldsymbol{v} \in \boldsymbol{H}_0(\mathrm{div};D) \mid (\nu\boldsymbol{v},\nabla\times\boldsymbol{\phi})_{\boldsymbol{L}^2(D)} = 0,\ \forall\boldsymbol{\phi} \in \boldsymbol{H}_0(\mathrm{curl};D)\}$ and accept as a fact that there is $\hat{C}_{\mathrm{PS}} > 0$ such that $\hat{C}_{\mathrm{PS}}\ell_D^{-1}\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \le \|\nabla\cdot\boldsymbol{v}\|_{L^2(D)}$ for all $\boldsymbol{v} \in \boldsymbol{Y}_{0\nu}$. Let $k \ge 0$ and consider the discrete space $\boldsymbol{Y}_{h0\nu} := \{\boldsymbol{v}_h \in \boldsymbol{P}^{\mathrm{d}}_{k,0}(\mathcal{T}_h) \mid (\nu\boldsymbol{v}_h,\nabla\times\boldsymbol{\phi}_h)_{\boldsymbol{L}^2(D)} = 0,\ \forall\boldsymbol{\phi}_h \in \boldsymbol{P}^{\mathrm{c}}_{k,0}(\mathcal{T}_h;\mathbb{C})\}$. Prove that there is $\hat{C}'_{\mathrm{PS}} > 0$ such that $\hat{C}'_{\mathrm{PS}}\|\boldsymbol{v}_h\|_{\boldsymbol{L}^2(D)} \le \ell_D\|\nabla\cdot\boldsymbol{v}_h\|_{L^2(D)}$ for all $\boldsymbol{v}_h \in \boldsymbol{Y}_{h0\nu}$. (*Hint*: adapt the proof of Theorem 44.6 using $\mathcal{J}^{\mathrm{d}}_{h0}$.)

Solution to exercises

Exercise 44.1 (Gradient). It is clear that $\nabla\phi \in \boldsymbol{H}(\mathrm{curl};D)$ for all $\phi \in H^1_0(D)$. Hence, we must just show that $(\nabla\phi)_{|\partial D}\times\boldsymbol{n} = \boldsymbol{0}$. Using the definition of γ^{c} in (4.11), we have

$$\langle\gamma^{\mathrm{c}}(\nabla\phi),\boldsymbol{l}\rangle_{\partial D} = \int_D \nabla\phi\cdot\nabla\times\boldsymbol{w}(\boldsymbol{l})\,\mathrm{d}x - \int_D (\nabla\times\nabla\phi)\cdot\boldsymbol{w}(\boldsymbol{l})\,\mathrm{d}x = \int_D \nabla\phi\cdot\nabla\times\boldsymbol{w}(\boldsymbol{l})\,\mathrm{d}x,$$

for all $\boldsymbol{l} \in \boldsymbol{H}^{\frac{1}{2}}(\partial D)$. Using the definition of γ^{d} in (4.12), we have

$$\langle \gamma^{\mathrm{c}}(\nabla\phi), \boldsymbol{l}\rangle_{\partial D} = -\int_D \phi \nabla\cdot(\nabla\times\boldsymbol{w}(\boldsymbol{l}))\,\mathrm{d}x + \langle \gamma^{\mathrm{d}}(\nabla\times\boldsymbol{w}(\boldsymbol{l})), \gamma^{\mathrm{g}}(\phi)\rangle_{\partial} = 0.$$

Hence, $\langle \gamma^{\mathrm{c}}(\nabla\phi), \boldsymbol{l}\rangle_{\partial} = 0$ for all $\boldsymbol{l} \in \boldsymbol{H}^{\frac{1}{2}}(\partial D)$. This proves that $\gamma^{\mathrm{c}}(\nabla\phi) = \boldsymbol{0}$.

Exercise 44.2 (Vector potential). The problem defining $\boldsymbol{z}_h \in \boldsymbol{X}_{h0\nu}$ such that

$$(\nabla\times\boldsymbol{z}_h, \nabla\times\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)} = (\nu\boldsymbol{v}, \boldsymbol{w}_h)_{\boldsymbol{L}^2(D)}, \qquad \forall \boldsymbol{w}_h \in \boldsymbol{X}_{h0\nu},$$

has a unique solution since the sesquilinear form is coercive and bounded on $\boldsymbol{X}_{h0\nu}$ (uniformly w.r.t. $h \in \mathcal{H}$) owing to Theorem 44.6. Moreover, the equality $(\nu\boldsymbol{v}, \boldsymbol{w}_h)_{\boldsymbol{L}^2(D)} = (\nabla\times\boldsymbol{z}_h, \nabla\times\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)}$ is valid for all $\boldsymbol{w}_h := \nabla m_h$ with $m_h \in M_{h0}$ owing to the assumption on $\boldsymbol{v}$ and the fact that $\nabla\times(\nabla m_h) = \boldsymbol{0}$. The conclusion follows from the identity $\boldsymbol{V}_{h0} = \boldsymbol{X}_{h0\nu} \oplus \nabla M_{h0}$.

Exercise 44.3 (Neumann condition). Let $\boldsymbol{x}_h \in \boldsymbol{X}_{h*\nu}$ be a nonzero discrete field. Let $\phi(\boldsymbol{x}_h) \in M_*$ be the solution to the following well-posed Neumann problem:

$$(\nu\nabla\phi(\boldsymbol{x}_h), \nabla m)_{\boldsymbol{L}^2(D)} = (\nu\boldsymbol{x}_h, \nabla m)_{\boldsymbol{L}^2(D)}, \qquad \forall m \in M_*.$$

Let $\boldsymbol{\xi}(\boldsymbol{x}_h) := \boldsymbol{x}_h - \nabla\phi(\boldsymbol{x}_h)$, so that $\boldsymbol{\xi}(\boldsymbol{x}_h) \in \boldsymbol{X}_{*\nu}$. Then we have

$$\boldsymbol{x}_h - \mathcal{J}_h^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)) = \mathcal{J}_h^{\mathrm{c}}(\boldsymbol{x}_h - \boldsymbol{\xi}(\boldsymbol{x}_h)) = \mathcal{J}_h^{\mathrm{c}}(\nabla(\phi(\boldsymbol{x}_h))) = \nabla(\mathcal{J}_h^{\mathrm{g}}(\phi(\boldsymbol{x}_h))),$$

where we used that $\mathcal{J}_h^{\mathrm{c}}(\boldsymbol{x}_h) = \boldsymbol{x}_h$ and the commuting properties of the quasi-interpolation operators $\mathcal{J}_h^{\mathrm{g}}$ and $\mathcal{J}_h^{\mathrm{c}}$. Since $\boldsymbol{x}_h \in \boldsymbol{X}_{h*\nu}$, we infer that $(\nu\boldsymbol{x}_h, \nabla(\mathcal{J}_h^{\mathrm{g}}(\phi(\boldsymbol{x}_h))))_{\boldsymbol{L}^2(D)} = 0$ (note that we can always shift $\mathcal{J}_h^{\mathrm{g}}(\phi(\boldsymbol{x}_h))$ by a constant without changing its gradient in such a way that this function is in M_*). We infer that

$$\begin{aligned}(\nu\boldsymbol{x}_h, \boldsymbol{x}_h)_{\boldsymbol{L}^2(D)} &= (\nu\boldsymbol{x}_h, \boldsymbol{x}_h - \mathcal{J}_h^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)} + (\nu\boldsymbol{x}_h, \mathcal{J}_h^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)}\\ &= (\nu\boldsymbol{x}_h, \mathcal{J}_h^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)}.\end{aligned}$$

Multiplying by $e^{\mathrm{i}\theta}$, taking the real part, and using the Cauchy–Schwarz inequality, we infer that

$$\nu_\flat \|\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)}^2 \le \nu_\sharp \|\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)} \|\mathcal{J}_h^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h))\|_{\boldsymbol{L}^2(D)}.$$

The uniform boundedness of $\mathcal{J}_h^{\mathrm{c}}$ on $\boldsymbol{L}^2(D)$ together with the Poincaré–Steklov inequality on $\boldsymbol{X}_{*\nu}$ implies that

$$\|\mathcal{J}_h^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h))\|_{\boldsymbol{L}^2(D)} \le \|\mathcal{J}_h^{\mathrm{c}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)} \|\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)} \le \|\mathcal{J}_h^{\mathrm{c}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)} \hat{C}_{\mathrm{PS}}^{-1} \ell_D \|\nabla\times\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)},$$

so that the expected result holds true with $\hat{C}'_{\mathrm{PS}} := \nu_{\sharp/\flat}^{-1} \|\mathcal{J}_h^{\mathrm{c}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}^{-1} \hat{C}_{\mathrm{PS}}$.

Exercise 44.4 (Discrete Poincaré–Steklov for $\nabla\cdot$). Let $\boldsymbol{x}_h \in \boldsymbol{Y}_{h0\nu}$ be a nonzero discrete field. Let $\boldsymbol{\zeta}(\boldsymbol{x}_h) \in \boldsymbol{M}_0 := \{\boldsymbol{v} \in \boldsymbol{H}_0(\mathrm{curl}; D) \mid \nabla\cdot\boldsymbol{v} = 0\}$ be the solution to the following well-posed problem (see (44.9)):

$$(\nu\nabla\times\boldsymbol{\zeta}(\boldsymbol{x}_h), \nabla\times\boldsymbol{m})_{\boldsymbol{L}^2(D)} = (\nu\boldsymbol{x}_h, \nabla\times\boldsymbol{m})_{\boldsymbol{L}^2(D)}, \qquad \forall \boldsymbol{m} \in \boldsymbol{M}_0.$$

Let us define $\boldsymbol{\xi}(\boldsymbol{x}_h) := \boldsymbol{x}_h - \nabla\times\boldsymbol{\zeta}(\boldsymbol{x}_h)$. This definition implies that $\boldsymbol{\xi}(\boldsymbol{x}_h) \in \boldsymbol{Y}_{0\nu}$. Indeed, any $\boldsymbol{\phi} \in \boldsymbol{H}_0(\mathrm{curl}; D)$ can be written as $\boldsymbol{\phi} := \boldsymbol{m} + \nabla\theta$ with $\boldsymbol{m} \in \boldsymbol{M}_0$ and $\theta \in H_0^1(D)$ owing to the Helmholtz decomposition from Lemma 44.1. Hence, we have

$$(\nu\boldsymbol{\xi}(\boldsymbol{x}_h), \nabla\times\boldsymbol{\phi})_{\boldsymbol{L}^2(D)} = (\nu\boldsymbol{\xi}(\boldsymbol{x}_h), \nabla\times\boldsymbol{m})_{\boldsymbol{L}^2(D)} = 0.$$

Invoking the quasi-interpolation operators $\mathcal{J}_{h0}^{\mathrm{c}}$ and $\mathcal{J}_{h0}^{\mathrm{d}}$ introduced in §23.3.3, we observe that

$$\boldsymbol{x}_h - \mathcal{J}_{h0}^{\mathrm{d}}(\boldsymbol{\xi}(\boldsymbol{x}_h)) = \mathcal{J}_{h0}^{\mathrm{d}}(\boldsymbol{x}_h - \boldsymbol{\xi}(\boldsymbol{x}_h)) = \mathcal{J}_{h0}^{\mathrm{d}}(\nabla\times(\boldsymbol{\zeta}(\boldsymbol{x}_h))) = \nabla\times(\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\zeta}(\boldsymbol{x}_h))),$$

where we used that $\mathcal{J}_{h0}^{\mathrm{d}}(\boldsymbol{x}_h) = \boldsymbol{x}_h$ and the commuting properties of the operators $\mathcal{J}_{h0}^{\mathrm{c}}$ and $\mathcal{J}_{h0}^{\mathrm{d}}$. Note that the above identity implies that $\nabla\cdot\boldsymbol{x}_h = \nabla\cdot\mathcal{J}_{h0}^{\mathrm{d}}(\boldsymbol{\xi}(\boldsymbol{x}_h))$. Since $\boldsymbol{x}_h \in \boldsymbol{Y}_{h0\nu}$, we infer that

$$\begin{aligned}(\nu\boldsymbol{x}_h, \boldsymbol{x}_h)_{\boldsymbol{L}^2(D)} &= (\nu\boldsymbol{x}_h, \boldsymbol{x}_h - \mathcal{J}_{h0}^{\mathrm{d}}(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)} + (\nu\boldsymbol{x}_h, \mathcal{J}_{h0}^{\mathrm{d}}(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)}\\ &= (\nu\boldsymbol{x}_h, \mathcal{J}_{h0}^{\mathrm{d}}(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)}.\end{aligned}$$

Multiplying by $e^{\mathrm{i}\theta}$, taking the real part, and using the Cauchy–Schwarz inequality, we infer that

$$\nu_\flat\|\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)}^2 \le \nu_\sharp\|\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)}\|\mathcal{J}_{h0}^{\mathrm{d}}(\boldsymbol{\xi}(\boldsymbol{x}_h))\|_{\boldsymbol{L}^2(D)}.$$

The uniform boundedness of $\mathcal{J}_{h0}^{\mathrm{d}}$ on $\boldsymbol{L}^2(D)$ together with the Poincaré–Steklov inequality for the divergence operator, that is,

$$\hat{C}_{\mathrm{PS}}\ell_D^{-1}\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \le \|\nabla\cdot\boldsymbol{v}\|_{\boldsymbol{L}^2(D)}, \qquad \forall\boldsymbol{v}\in\boldsymbol{Y}_{0\nu},$$

imply that

$$\begin{aligned}\|\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h))\|_{\boldsymbol{L}^2(D)} &\le \|\mathcal{J}_{h0}^{\mathrm{d}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}\|\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)}\\ &\le \|\mathcal{J}_{h0}^{\mathrm{d}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}\hat{C}_{\mathrm{PS}}^{-1}\ell_D\|\nabla\cdot\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)}\\ &\le \|\mathcal{J}_{h0}^{\mathrm{d}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}\hat{C}_{\mathrm{PS}}^{-1}\ell_D\|\nabla\cdot\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)},\end{aligned}$$

so that the expected discrete Poincaré–Steklov inequality holds true with $\hat{C}'_{\mathrm{PS}} := \nu_{\sharp/\flat}^{-1}\|\mathcal{J}_{h0}^{\mathrm{d}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}^{-1}\hat{C}_{\mathrm{PS}}$.

Chapter 45

Maxwell's equations: further topics

Exercises

Exercise 45.1 (Identity for $n_\sharp$). Prove (45.13b). (*Hint*: use the mollification operators $\mathcal{K}_\delta^{\mathrm{c}} : \boldsymbol{L}^1(D) \to \boldsymbol{C}^\infty(\overline{D})$ and $\mathcal{K}_\delta^{\mathrm{d}} : \boldsymbol{L}^1(D) \to \boldsymbol{C}^\infty(\overline{D})$ from §23.1, and adapt the proof of Lemma 40.5.)

Exercise 45.2 (Consistency/boundedness). Prove Lemma 45.5. (*Hint*: adapt the proof of Lemma 41.7 and use Lemma 45.4.)

Exercise 45.3 (Least-squares penalty on divergence). (i) Prove Proposition 45.10. (*Hint*: use Lemma 44.1 to write $\boldsymbol{A} := \boldsymbol{A}_0 + \nabla p$, where $\boldsymbol{A}_0$ is divergence-free and $p \in H_0^1(D)$, and prove that $p = 0$.) (ii) Prove (45.22). (*Hint*: use Lemma 44.4 for $\boldsymbol{A} - \nabla p$.)

Solution to exercises

Exercise 45.1 (Identity for $n_\sharp$). Let us set

$$n_{\sharp\delta}(\boldsymbol{a}, \boldsymbol{b}_h) := \sum_{F \in \mathcal{F}_h^\partial} \langle (\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{\sigma}(\boldsymbol{a}))_{|K_l} \times \boldsymbol{n})_{|F}, \Pi_F(\boldsymbol{b}_h) \rangle_F.$$

Owing to (45.9) and the commuting property $\nabla\times(\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{v})) = \mathcal{K}_\delta^{\mathrm{d}}(\nabla\times\boldsymbol{v})$, we have

$$n_{\sharp\delta}(\boldsymbol{a}, \boldsymbol{b}_h) = \sum_{F \in \mathcal{F}_h^\partial} \int_{K_l} \left(\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{\sigma}(\boldsymbol{a}))\cdot\nabla\times L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h)) - \mathcal{K}_\delta^{\mathrm{d}}(\nabla\times\boldsymbol{\sigma}(\boldsymbol{a}))\cdot L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h)) \right) \mathrm{d}x,$$

and owing to Theorem 23.4, we infer that

$$\lim_{\delta\to 0} n_{\sharp\delta}(\boldsymbol{a}, \boldsymbol{b}_h) = \sum_{F \in \mathcal{F}_h^\partial} \int_{K_l} \Big(\boldsymbol{\sigma}(\boldsymbol{a})\cdot\nabla\times L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h)) - \nabla\times\boldsymbol{\sigma}(\boldsymbol{a})\cdot L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h)) \Big) \mathrm{d}x = n_\sharp(\boldsymbol{a}, \boldsymbol{b}_h).$$

Since $\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{\sigma}(\boldsymbol{a}))$ is smooth, we also have

$$\begin{aligned}
n_{\sharp\delta}(\boldsymbol{a},\boldsymbol{b}_h) &= \sum_{F\in\mathcal{F}_h^\partial}\int_{\partial K_l}(\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{\sigma}(\boldsymbol{a}))\times\boldsymbol{n}_{K_l})\cdot L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h))\,\mathrm{d}s\\
&= \sum_{F\in\mathcal{F}_h^\partial}\int_{F}(\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{\sigma}(\boldsymbol{a}))\times\boldsymbol{n})\cdot\Pi_F(\overline{\boldsymbol{b}}_h)\,\mathrm{d}s\\
&= \int_{\partial D}(\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{\sigma}(\boldsymbol{a}))\times\boldsymbol{n})\cdot\overline{\boldsymbol{b}}_h\,\mathrm{d}s\\
&= \int_D\Big(\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{\sigma}(\boldsymbol{a}))\cdot\nabla\times\overline{\boldsymbol{b}}_h-\nabla\times(\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{\sigma}(\boldsymbol{a})))\cdot\overline{\boldsymbol{b}}_h\Big)\,\mathrm{d}x\\
&= \int_D\Big(\mathcal{K}_\delta^{\mathrm{c}}(\boldsymbol{\sigma}(\boldsymbol{a}))\cdot\nabla\times\overline{\boldsymbol{b}}_h-\mathcal{K}_\delta^{\mathrm{d}}(\nabla\times(\boldsymbol{\sigma}(\boldsymbol{a})))\cdot\overline{\boldsymbol{b}}_h\Big)\,\mathrm{d}x.
\end{aligned}$$

We conclude by passing to the limit $\delta\to 0$ on the right-hand side and using again Theorem 23.4.

Exercise 45.2 (Consistency/boundedness). Let $\boldsymbol{a}_h,\boldsymbol{b}_h\in\boldsymbol{V}_h$. Using the identities from Lemma 45.4, we infer that

$$\begin{aligned}
\langle\delta_h(\boldsymbol{a}_h),\boldsymbol{b}_h\rangle_{\boldsymbol{V}_h',\boldsymbol{V}_h} &= a_{\nu,\kappa}(\boldsymbol{\theta}_h,\boldsymbol{b}_h)-n_\sharp(\boldsymbol{\theta}_h,\boldsymbol{b}_h)-\sum_{F\in\mathcal{F}_h^\partial}\eta_0 e^{-\mathrm{i}\theta}\frac{|\kappa_{K_l}|^2}{\kappa_{r,K_l}h_F}\int_F(\boldsymbol{a}_h\times\boldsymbol{n})\cdot(\overline{\boldsymbol{b}}_h\times\boldsymbol{n})\,\mathrm{d}s\\
&=: \mathfrak{T}_1+\mathfrak{T}_2+\mathfrak{T}_3,
\end{aligned}$$

with $\boldsymbol{\theta}_h:=\boldsymbol{A}-\boldsymbol{a}_h$. The Cauchy–Schwarz inequality implies that

$$\begin{aligned}
|\mathfrak{T}_1| \le{}& \left(\sum_{K\in\mathcal{T}_h}\left(\frac{|\nu_K|^2}{\nu_{r,K}}\|\boldsymbol{\theta}_h\|_{\boldsymbol{L}^2(K)}^2+\frac{|\kappa_K|^2}{\kappa_{r,K}}\|\nabla\times\boldsymbol{\theta}_h\|_{\boldsymbol{L}^2(K)}^2\right)\right)^{\frac12}\\
&\times\left(\sum_{K\in\mathcal{T}_h}\left(\nu_{r,K}\|\boldsymbol{b}_h\|_{\boldsymbol{L}^2(K)}^2+\kappa_{r,K}\|\nabla\times\boldsymbol{b}_h\|_{\boldsymbol{L}^2(K)}^2\right)\right)^{\frac12}.
\end{aligned}$$

Hence, $|\mathfrak{T}_1|\le\|\boldsymbol{\theta}_h\|_{\boldsymbol{V}_\sharp}\|\boldsymbol{b}_h\|_{\boldsymbol{V}_h}$. Moreover, recalling that $\lambda_F^{-1}:=\frac{\kappa_{r,K_l}}{|\kappa_{K_l}|^2}$, the boundedness estimate (45.14) on $n_\sharp$ yields

$$|\mathfrak{T}_2|\le c\left(\sum_{F\in\mathcal{F}_h^\partial}\frac{\kappa_{r,K_l}}{|\kappa_{K_l}|^2}h_{K_l}^{2d(\frac12-\frac1p)}\|\boldsymbol{\sigma}(\boldsymbol{\theta}_h)\|_{\boldsymbol{V}^{\mathrm{c}}(K_l)}^2\right)^{\frac12}|\boldsymbol{b}_h|_\partial.$$

Since κ is constant on K_l, we have

$$\|\boldsymbol{\sigma}(\boldsymbol{\theta}_h)\|_{\boldsymbol{V}^{\mathrm{c}}(K_l)}\le|\kappa_{K_l}|\Big(\|\nabla\times\boldsymbol{\theta}_h\|_{\boldsymbol{L}^p(K)}+h_K^{1+d(\frac1p-\frac1q)}\|\nabla\times(\nabla\times\boldsymbol{\theta}_h)\|_{\boldsymbol{L}^q(K)}\Big).$$

Hence, $|\mathfrak{T}_2|\le c\|\boldsymbol{\theta}_h\|_{\boldsymbol{V}_\sharp}\|\boldsymbol{b}_h\|_{\boldsymbol{V}_h}$. Finally, the Cauchy–Schwarz inequality leads to $|\mathfrak{T}_3|\le|\boldsymbol{a}_h|_\partial|\boldsymbol{b}_h|_\partial$, and we have $|\boldsymbol{\theta}_h|_\partial=|\boldsymbol{a}_h|_\partial$ since $\boldsymbol{A}_{|\partial D}\times\boldsymbol{n}=\boldsymbol{0}$.

Exercise 45.3 (Least-squares penalty on divergence). (i) Assume that $\boldsymbol{A}\in\boldsymbol{H}_0(\mathrm{curl};D)$ solves (44.1). We have already established that $\nabla\cdot\boldsymbol{A}=0$. Hence, $\boldsymbol{A}\in\boldsymbol{Z}_0$ and $a_{\nu,\kappa,\eta}(\boldsymbol{A},\boldsymbol{b})=a_{\nu,\kappa}(\boldsymbol{A},\boldsymbol{b})$ for all $\boldsymbol{b}\in\boldsymbol{Z}_0$. This shows that $\boldsymbol{A}$ solves (45.20). Conversely, assume that $\boldsymbol{A}$ solves (45.20).

Recalling Lemma 44.1, let $p \in H^1_0(D)$ be s.t. $\boldsymbol{A} := \boldsymbol{A}_0 + \nabla p$, where $\boldsymbol{A}_0$ is divergence-free, i.e., $(\nabla p, \nabla q)_{\boldsymbol{L}^2(D)} = (\boldsymbol{A}, \nabla q)_{\boldsymbol{L}^2(D)}$ for all $q \in H^1_0(D)$. Using the test function $\boldsymbol{b} = e^{-\mathrm{i}\theta}\nabla p$ in (45.20), observing that $\nabla\times(\nabla p) = \boldsymbol{0}$, $(\boldsymbol{A}, \nabla p)_{\boldsymbol{L}^2(D)} = \|\nabla p\|^2_{\boldsymbol{L}^2(D)}$, $\Delta p = \nabla\cdot\boldsymbol{A} \in L^2(D)$, and $(\boldsymbol{f}, \nabla p)_{\boldsymbol{L}^2(D)} = 0$, and taking the real part leads to

$$\nu_\flat \|\nabla p\|^2_{\boldsymbol{L}^2(D)} + \eta\kappa_\flat \|\Delta p\|^2_{L^2(D)} \le 0,$$

whence we infer that $p = 0$. Hence, $\boldsymbol{A} = \boldsymbol{A}_0$ is divergence-free. Finally, we have $a_{\nu,\kappa,\eta}(\boldsymbol{A}, \boldsymbol{b}) = a_{\nu,\kappa}(\boldsymbol{A}, \boldsymbol{b})$ for all $\boldsymbol{b} \in \boldsymbol{Z}_0$, and this equality can be extended to any $\boldsymbol{b} \in \boldsymbol{H}_0(\mathrm{curl}; D)$ by density of smooth functions in $\boldsymbol{H}_0(\mathrm{curl}; D)$. We have thus proved that $\boldsymbol{A}$ solves (44.1).
(ii) Since $\nabla\cdot\boldsymbol{A}_0$ is divergence-free, $\boldsymbol{A}_0 \in \boldsymbol{H}_0(\mathrm{curl}; D)$, and $\nabla\times\boldsymbol{A}_0 = \nabla\times\boldsymbol{A}$, Lemma 44.4 implies that

$$\hat{C}_{\mathrm{PS}}\ell_D^{-1}\|\boldsymbol{A} - \nabla p\|_{\boldsymbol{L}^2(D)} = \hat{C}_{\mathrm{PS}}\ell_D^{-1}\|\boldsymbol{A}_0\|_{\boldsymbol{L}^2(D)} \le \|\nabla\times\boldsymbol{A}_0\|_{\boldsymbol{L}^2(D)} = \|\nabla\times\boldsymbol{A}\|_{\boldsymbol{L}^2(D)}.$$

Invoking the triangle inequality and since $\|\nabla p\|_{\boldsymbol{L}^2(D)} \le C_{\mathrm{PS}}^{-1}\ell_D\|\nabla\cdot\boldsymbol{A}\|_{L^2(D)}$, where C_{PS} is s.t. $C_{\mathrm{PS}}\ell_D^{-1}\|q\|_{L^2(D)} \le \|\nabla q\|_{\boldsymbol{L}^2(D)}$ for all $q \in H^1_0(D)$, we conclude that (45.22) holds true with $\hat{C}''_{\mathrm{PS}} := \min(\hat{C}_{\mathrm{PS}}, C_{\mathrm{PS}})$.

Chapter 46

Symmetric elliptic eigenvalue problems

Exercises

Exercise 46.1 (Spectrum). Let L be a complex Banach space. Let $T \in \mathcal{L}(L)$. (i) Show that $(\lambda T)^* = \overline{\lambda} T^*$ for all $\lambda \in \mathbb{C}$. (ii) Show that $\sigma_{\mathrm{r}}(T) \subset \mathrm{conj}(\sigma_{\mathrm{p}}(T^*)) \subset \sigma_{\mathrm{r}}(T) \cup \sigma_{\mathrm{p}}(T)$. (*Hint*: use Corollary C.15.) (iii) Show that the spectral radius of T verifies $r(T) \le \limsup_{n\to\infty} \|T^n\|_{\mathcal{L}(L)}^{\frac{1}{n}}$. (*Hint*: consider $\sum_{n\in\mathbb{N}} (\mu^{-1}T)^n$ and use the root test: the complex-valued series $\sum_{n\in\mathbb{N}} a_n$ converges absolutely if $\limsup_{n\to\infty} |a_n|^{\frac{1}{n}} < 1$.)

Exercise 46.2 (Ascent, algebraic and geometric multiplicities). (i) Let $T \in \mathcal{L}(L)$. Let μ be an eigenvalue of T and let $K_i := \ker(\mu I_L - T)^i$ for all $i \in \mathbb{N}\backslash\{0\}$. Prove that $K_1 \subset K_2 \ldots$, and assuming that there is $j \ge 1$ s.t. $K_j = K_{j+1}$, show that $K_j = K_{j'}$ for all $j' > j$. (ii) Assume that μ has a finite ascent α, and finite algebraic multiplicity m and geometric multiplicity g. Show that $\alpha + g - 1 \le m \le \alpha g$. (*Hint*: letting $g_i := \dim(K_i)$ for all $i \in \{1{:}\alpha\}$, prove that $g_1 + i - 1 \le g_i$ and $g_i \le g_{i-1} + g_1$.) (iii) Compute the ascent, algebraic multiplicity, and geometric multiplicity of the eigenvalues of following matrices and verify the two inequalities from Step (i):

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Exercise 46.3 (Eigenspaces). The following three questions are independent. (i) Suppose $V = V_1 \oplus V_2$ and consider $T \in \mathcal{L}(V)$ defined by $T(v_1 + v_2) := v_1$ for all $v_1 \in V_1$ and all $v_2 \in V_2$. Find all the eigenvalues and eigenspaces of T. (ii) Let $T \in \mathcal{L}(V)$. Assume that S is invertible. Prove that $S^{-1}TS$ and T have the same eigenvalues. What is the relationship between the eigenvectors of T and those of $S^{-1}TS$? (iii) Let V be a finite-dimensional vector space. Let $\{v_n\}_{n\in\{1:m\}} \subset V$, $m \ge 1$. Show that the vectors $\{v_n\}_{n\in\{1:m\}}$ are linearly independent iff there exists $T \in \mathcal{L}(V)$ such that $\{v_n\}_{n\in\{1:m\}}$ are eigenvectors of T corresponding to distinct eigenvalues.

Exercise 46.4 (Volterra operator). Let $L := L^2((0,1);\mathbb{C})$ and let $T : L \to L$ be s.t. $T(f)(x) := \int_0^x f(t)\,\mathrm{d}t$ for a.e. $x \in (0,1)$. Notice that T is a Hilbert–Schmidt operator, but this exercise is

meant to be done without using this fact. (i) Show that $T^{\mathsf{H}}(g) = \int_x^1 g(t)\,\mathrm{d}t$ for all $g \in L^2((0,1);\mathbb{C})$. (ii) Show that T is injective. (*Hint*: use Theorem 1.32.) (iii) Show that $0 \in \sigma_{\mathrm{c}}(T)$. (iv) Show that $\sigma_{\mathrm{p}}(T) = \emptyset$. (v) Prove that $\mu I_L - T$ is bijective if $\mu \neq 0$. (vi) Determine $\rho(T)$, $\sigma_{\mathrm{p}}(T)$, $\sigma_{\mathrm{c}}(T)$, $\sigma_{\mathrm{r}}(T)$. Do the same for T^{H}.

Exercise 46.5 (Riesz–Fréchet). Let H be a finite-dimensional complex Hilbert space with orthonormal basis $\{e_i\}_{i\in\{1:n\}}$ and inner product $(\cdot,\cdot)_H$. (i) Let g be an antilinear form on H, i.e., $g \in H'$. Show that $(J_H^{\mathrm{RF}})^{-1}(g) = \sum_{i\in\{1:n\}} g(e_i)e_i$ with $g(e_i) := \langle g, e_i\rangle_{H',H}$, $\forall i \in \{1{:}n\}$. Is $(J_H^{\mathrm{RF}})^{-1} : H' \to H$ linear or antilinear? (ii) Let g be a linear form on H. Show that $x_g := \sum_{i\in\{1:n\}} \overline{g(e_i)}e_i$ is s.t. $\langle g, y\rangle_{H',H} = \overline{(x_g, y)_H}$. Is the map $H' \ni g \mapsto x_g \in H$ linear or antilinear?

Exercise 46.6 (Symmetric operator). Let L be a complex Hilbert space and $T \in \mathcal{L}(L)$ be a symmetric operator. (i) Show that $\sigma(T) \subset \mathbb{R}$. (*Hint*: compute $\Im((T(v) - \mu v, v)_L$ and show that $|\Im(\mu)|\|v\|_L^2 \le |(T(v) - \mu v, v)_L|$ for all $v \in L$.) (ii) Prove that $\sigma_{\mathrm{r}}(T) = \emptyset$. (*Hint*: apply Corollary C.15.) (iii) Show that the ascent of each $\mu \in \sigma_{\mathrm{p}}(T)$ is equal to 1. (*Hint*: compute $\|(\mu I_L - T)(x)\|_L^2$ with $x \in \ker(\mu I_L - T)^2$.)

Exercise 46.7 ($H^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is not compact). (i) Let $\chi(x) := 1 + x$ if $-1 \le x \le 0$, $\chi(x) := 1 - x$ if $0 \le x \le 1$ and $\chi(x) := 0$ if $|x| \ge 1$. Show that $\chi \in H^1(\mathbb{R})$. (ii) Let $v_n(x) := \chi(x-n)$ for all $n \in \mathbb{N}$. Show that $(v_n)_{n\in\mathbb{N}}$ converges weakly to 0 in $L^2(\mathbb{R})$ (see Definition C.28). (iii) Show that the embedding $H^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is not compact. (*Hint*: argue by contradiction using Theorem C.23.)

Exercise 46.8 ($B^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact). (i) Show that the embedding $B^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact, where $B^1(\mathbb{R}) := \{v \in H^1(\mathbb{R}) \mid xv \in L^2(\mathbb{R})\}$. (*Hint*: let $(u_n)_{n\in\mathbb{N}}$ be a bounded sequence in $B^1(\mathbb{R})$, build nested subsets $J_k \subset \mathbb{N}$, $\forall k \in \mathbb{N}\backslash\{0\}$, s.t. the sequence $(u_{n|(-k,k)})_{n\in J_k}$ converges in $L^2(-k,k)$.) (ii) Give a sufficient condition on $\alpha \in \mathbb{R}$ so that $B^1_\alpha(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact, where $B^1_\alpha(\mathbb{R}) := \{v \in H^1(\mathbb{R}) \mid |x|^\alpha v \in L^2(\mathbb{R})\}$.

Exercise 46.9 (Hausdorff–Toeplitz theorem). The goal of this exercise is to prove that the numerical range of a bounded linear operator in a Hilbert space is convex; see also Gustafson [23]. Let L be a complex Hilbert space and let $S_L(1) := \{x \in L \mid \|x\|_L = 1\}$ be the unit sphere in L. Let $T \in \mathcal{L}(L)$ and let $W(T) := \{\alpha \in \mathbb{C} \mid \exists x \in S_L(1),\ \alpha = (T(x), x)_L\}$ be the numerical range of T. Let $\gamma, \mu \in W(T)$, $\gamma \neq \mu$, and $x_1, x_2 \in S_L(1)$ be s.t. $(T(x_1), x_1)_L = \gamma$, $(T(x_2), x_2)_L = \mu$. Let $T' := \frac{1}{\mu-\gamma}(T - \gamma I_L)$. (i) Compute $(T'(x_1), x_1)_L$ and $(T'(x_2), x_2)_L$. (ii) Prove that there exists $\theta \in [0, 2\pi)$ s.t. $\Im(e^{\mathrm{i}\theta}(T'(x_1), x_2)_L + e^{-\mathrm{i}\theta}(T'(x_2), x_1)_L) = 0$. (iii) Let $x_1' := e^{\mathrm{i}\theta}x_1$. Compute $(T'(x_1'), x_1')_L$. (iv) Let $\lambda \in [0,1]$. Show that the following problem has at least one solution: Find $\alpha, \beta \in \mathbb{R}$ s.t. $\|\alpha x_1' + \beta x_2\|_L = 1$ and $(T'(\alpha x_1' + \beta x_2), \alpha x_1' + \beta x_2)_L = \lambda$. (*Hint*: view the two equations as those of an ellipse and an hyperbola, respectively, and determine how these curves cross the axes.) (v) Prove that $W(T)$ is convex. (*Hint*: compute $(T(\alpha x_1' + \beta x_2), \alpha x_1' + \beta x_2)_L$.)

Solution to exercises

Exercise 46.1 (Spectrum). (i) Recalling that, by convention, L' is composed of antilinear forms, we have

$$\langle (\lambda T)^*(l'), l\rangle_{L',L} = \langle l', \lambda T(l)\rangle_{L',L} = \overline{\lambda}\langle l', T(l)\rangle_{L',L} = \overline{\lambda}\langle T^*(l'), l\rangle_{L',L},$$

for all $\lambda \in \mathbb{C}$, all $l \in L$, and all $l' \in L'$. Hence, we have

$$(\lambda T)^*(l') = \overline{\lambda}T^*(l'), \qquad \forall l' \in L'.$$

This proves that $(\lambda T)^* = \overline{\lambda} T^*$ for all $\lambda \in \mathbb{C}$.
(ii) Let us start by showing that $\sigma_{\rm r}(T) \subset {\rm conj}(\sigma_{\rm p}(T^*))$. Let $\mu \in \sigma_{\rm r}(T)$. Then $\mu I_L - T$ is injective and $\overline{{\rm im}(\mu I_L - T)} \neq L$, i.e., ${\rm im}(\mu I_L - T)$ is not dense in L. Using a corollary of Hahn–Banach's theorem (Corollary C.15), we infer that there exists $0 \neq l' \in L'$ such that $0 = \langle l', (\mu I_L - T)(x)\rangle_{L',L} = \langle \overline{\mu} l' - T^*(l'), x\rangle_{L',L}$ for all $x \in L$ (recall that $(\mu I_L)^* = \overline{\mu} I_L^*$). This means that $\overline{\mu} l' - T^*(l') = 0$, i.e., $\overline{\mu} I_{L'} - T^*$ is not injective. This proves that $\overline{\mu}$ is an eigenvalue of T^*, i.e., $\sigma_{\rm r}(T) \subset {\rm conj}(\sigma_{\rm p}(T^*))$. We now prove the second inclusion. Let $\mu \in {\rm conj}(\sigma_{\rm p}(T^*))$ and $0 \neq l' \in \ker(\overline{\mu} I_L - T^*)$. Then $0 = \langle \overline{\mu} l' - T^*(l'), l\rangle_{L',L} = \langle l', (\mu I_L - T)(l)\rangle_{L',L}$ for all $l \in L$. Hence, ${\rm im}(\mu I_L - T)$ is not dense in L. This means that $\mu \in \sigma(T)$. But $\mu \notin \sigma_{\rm c}(T)$. Hence, $\mu \in \sigma_{\rm p}(T) \cup \sigma_{\rm r}(T)$.
(iii) Let us set $\tilde{r}(T) := \limsup_{n\to\infty} \|T^n\|_{\mathcal{L}(L)}^{\frac{1}{n}}$. Let $\mu \in \mathbb{C}$ be s.t. $|\mu| > \tilde{r}(T)$ (notice that $\mu \neq 0$). We have to show that $\mu I_L - T$ is bijective, which is equivalent to show that $I_L - \mu^{-1} T$ is bijective. The root test shows that the series $\sum_{k\in\mathbb{N}} \|(\mu^{-1}T)^k\|_{\mathcal{L}(L)}$ is convergent. It follows that the sequence $S_n := \sum_{k\in\{0:n\}} (\mu^{-1}T)^k$ is Cauchy in $\mathcal{L}(L)$. Since $\mathcal{L}(L)$ is complete, there is S s.t. $S_n \to S$ in $\mathcal{L}(L)$. But

$$\begin{aligned}(I_L - \mu^{-1}T)S_n &= (I_L - \mu^{-1}T) \sum_{k\in\{0:n\}} (\mu^{-1}T)^k \\ &= \sum_{k\in\{0:n\}} \mu^{-1}T^k - \sum_{k\in\{1:n+1\}} (\mu^{-1}T)^k \\ &= I_L - (\mu^{-1}T)^{n+1}.\end{aligned}$$

Notice that $\lim_{n\to\infty} \|(\mu^{-1}T)^n\|_{\mathcal{L}(L)} = 0$ since the series $\sum_{k\in\mathbb{N}} \|(\mu^{-1}T)^k\|_{\mathcal{L}(L)}$ is convergent. In conclusion, we have

$$(I_L - \mu^{-1}T)S = \lim_{n\to\infty} (I_L - \mu^{-1}T)S_n = I_L,$$

which proves that $I_L - \mu^{-1}T$ is invertible. Hence, $\mu \in \rho(T)$, which, in turn, proves that $r(T) \le \tilde{r}(T)$.

Exercise 46.2 (Ascent, algebraic and geometric multiplicities). (i) Let $K_i := \ker(\mu I_L - T)^i$ for all $i \in \mathbb{N} \setminus \{0\}$. Let $x \in K_i$. We have

$$(\mu I_L - T)^{i+1}(x) = ((\mu I_L - T) \circ (\mu I_L - T)^i)(x) = (\mu I_L - T)(0) = 0,$$

showing that $x \in K_{i+1}$. Hence, $K_i \subset K_{i+1}$ for all $i \ge 1$. Assume now that $K_j = K_{j+1}$ for some $j \ge 1$. Let us show by an induction argument on p that $K_j = K_{j+p}$. The statement holds true for $p = 1$. Assume that it holds true for some $p \ge 1$ and let us show that $K_j = K_{j+p+1}$. Since $K_j = K_{j+p} \subset K_{j+p+1}$, it suffices to show that $K_{j+p+1} \subset K_{j+p}$. Let $x \in K_{j+p+1}$. Since $(\mu I_L - T)^{j+p+1}(x) = 0$, we have $(\mu I_L - T)(x) \in K_{j+p} = K_{j+p-1}$ by the induction assumption, so that $x \in K_{j+p}$. This completes the proof.
(ii) Let $g_i := \dim(K_i)$ for all $i \in \{1{:}\alpha\}$. The definition of the ascent implies that $K_1 \subsetneq K_2 \subsetneq \ldots \subsetneq K_\alpha = K_{\alpha+j}$ for all $j \in \mathbb{N}$. As a result, we have $g_1 + i - 1 \le g_i$ for all $i \in \{1{:}\alpha\}$. Since $g_1 = g$ and $g_\alpha = m$, this implies that $\alpha + g - 1 \le m$.
Next, let us prove that $g_i \le g_{i-1} + g_1$ for all $i \in \{1{:}\alpha\}$. Once this is established, it follows that $g_2 \le 2g_1 = 2g$, $g_3 \le g_2 + g_1 \le 3g$, and so on, so that $m = g_\alpha \le \alpha g$. We start by writing $K_i = K_{i-1} \oplus Y_i$ (this is legitimate since we are working with finite-dimensional spaces). Let us show that $k := \dim(Y_i) \le g := g_1$. Notice that $k \ge 1$ since $i \in \{1{:}\alpha\}$. Let $(u_j)_{j\in\{1:k\}}$ be a basis of Y_i and let us verify that the vectors $((\mu I_L - T)^{i-1}(u_j))_{j\in\{1:k\}}$ are linearly independent. Let $(\alpha_j)_{j\in\{1:k\}}$ be k scalars such that $\sum_{j\in\{1:k\}} \alpha_j (\mu I_L - T)^{i-1}(u_j) = 0$. We have $\sum_{j\in\{1:k\}} \alpha_j u_j \in K_{i-1}$. But $\sum_{j\in\{1:k\}} \alpha_j u_j \in Y_i$ by definition. Since $K_{i-1} \cap Y_i = \{0\}$, we must have $\sum_{j\in\{1:k\}} \alpha_j u_j = 0$. Since $(u_j)_{j\in\{1:k\}}$ is a basis of Y_i, it follows that $\alpha_j = 0$ for all $j \in \{1{:}k\}$. This proves that indeed the

vectors $((\mu I_L - T)^{i-1}(u_j))_{j\in\{1:k\}}$ are linearly independent. But these vectors are also members of K_1 since $u_j \in K_i$ for each $j \in \{1{:}k\}$. This proves that $k \le g$. In conclusion, we have shown that $g_i \le g_{i-1} + g_1$.
(iii) Let $T : \mathbb{R}^4 \to \mathbb{R}^4$ be the operator defined by $T(X) := \mathbb{A}X$ for all $X \in L := \mathbb{R}^4$, with

$$\mathbb{A} := \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Hence, 1 is the only eigenvalue of T. A direct computation shows that

$$\mathbb{I}_4 - \mathbb{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (\mathbb{I}_4 - \mathbb{A})^2 = \begin{pmatrix} 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (\mathbb{I}_4 - \mathbb{A})^3 = \mathbb{O}_4.$$

Hence, $\ker(I_L - T)^2 \neq \ker(I_L - T)^3$, but $\ker(I_L - T)^3 = \ker(I_L - T)^4 = \mathbb{R}^4$. Thus, the ascent of $\mu = 1$ is $\alpha = 3$. Moreover, $\dim(\ker(I_L - T)^3) = 4$ and $\dim(\ker(I_L - T)) = 2$, i.e., the algebraic multiplicity is $m = 4$ and the geometric multiplicity is $g = 2$. Notice that we have $\alpha + g - 1 = 4 = m \le 6 = \alpha g$.
Let now $T : \mathbb{R}^4 \to \mathbb{R}^4$ be the operator defined by $T(X) := \mathbb{A}X$ for all $X \in L := \mathbb{R}^4$, with

$$A := \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Hence, 1 is the only eigenvalue of T. A direct computation shows that

$$\mathbb{I}_4 - \mathbb{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (\mathbb{I}_4 - \mathbb{A})^2 = \mathbb{O}_4.$$

Thus, $\ker(I_L - T) \neq \ker(I_L - T)^2$, but $\ker(I_L - T)^2 = \ker(I_L - T)^3 = \mathbb{R}^4$. This shows that the ascent of $\mu = 1$ is $\alpha = 2$. Moreover, $\dim(\ker(I_L - T)^2) = 4$ and $\dim(\ker(I_L - T)) = 2$, i.e., the algebraic multiplicity is $m = 4$ and the geometric multiplicity is $g = 2$. Notice that $\alpha + g - 1 = 3 \le 4 = m = \alpha g$.
Let finally $T : \mathbb{R}^4 \to \mathbb{R}^4$ be the operator defined by $T(X) := \mathbb{A}X$ for all $X \in L := \mathbb{R}^4$, with

$$\mathbb{A} := \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Hence, 1 is the only eigenvalue of T. A direct computation shows that

$$\mathbb{I}_4 - \mathbb{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (\mathbb{I}_4 - \mathbb{A})^2 = \mathbb{O}_4.$$

Thus, $\ker(I_L - T) \neq \ker(I_L - T)^2$, but $\ker(I_L - T)^2 = \ker(I_L - T)^3 = \mathbb{R}^4$. This shows that the ascent of $\mu = 1$ is $\alpha = 2$. Moreover, $\dim(\ker(I_L - T)^2) = 4$ and $\dim(\ker(I_L - T)) = 3$, i.e., the algebraic multiplicity is $m = 4$ and the geometric multiplicity is $g = 3$. Notice that $\alpha + g - 1 = 4 = m \leq 6 = \alpha g$.

Exercise 46.3 (Eigenspaces). (i) Let $\mu \in \mathbb{C}$ and let $v := v_1 + v_2 \in V_1 \oplus V_2$. Then (μ, v) is an eigenpair iff $\mu v = T(v) = v_1$. Hence, $(\mu - 1)v_1 + \mu v_2 = 0$. But the sum $V_1 \oplus V_2$ being direct, we infer that $(\mu - 1)v_1 = 0$ and $\mu v_2 = 0$. If $\mu = 1$, then $v_2 = 0$. If $\mu = 0$, then $v_1 = 0$. If $\mu \notin \{0, 1\}$, then $v_1 = 0$ and $v_2 = 0$. In conclusion, the eigenvalues are $\{0, 1\}$, and the associated eigenspaces are $\ker(T) = V_2$ and $\ker(I_L - T) = V_1$.
(ii) Let $S, T \in \mathcal{L}(V)$ with S invertible. Let $\mu \in \mathbb{C}$, $v \in V$. Then $T(v) = \mu v$ iff $S^{-1}TSS^{-1}v = \mu S^{-1}v$. This shows that $S^{-1}\ker(\mu I_L - T) = \ker(\mu I_L - S^{-1}TS)$. Thus, the eigenvalues of T and $S^{-1}TS$ are identical, and v is an eigenvector of T iff $S^{-1}v$ is an eigenvector of $S^{-1}TS$.
(iii) Let V be a finite-dimensional vector space. Let $v_1, \dots, v_m \in V$ and let us assume that $v_1, \dots, v_m$ are linearly independent. If $m < \dim(V) = n$, let $\{v_{m+1}, \dots, v_n\}$ be vectors that make $\{v_1, \dots, v_n\}$ a basis of V. Let $T : V \to V$ be defined by $Tv_i := iv_i$ for all $i \in \{1{:}n\}$. Then $\{v_1, \dots, v_n\}$ is a basis of eigenvectors of T, and the eigenvalues are $\{1, \dots, n\}$. Conversely, assume that there exists $T \in \mathcal{L}(V)$ such that $v_1, \dots, v_m$ are eigenvectors of T corresponding to distinct eigenvalues, $\mu_1, \dots, \mu_m$. Assume that $v_1, \dots, v_m$ are linearly dependent. Without loss of generality, let us assume that v_1 depends on $(v_l)_{l \in L}$ where $L \subset \{2, \dots, m\}$, and the vectors $(v_l)_{l \in L}$ are linearly independent. Then $v_1 = \sum_{l \in L} \alpha_l v_l$ and

$$T(v_1) = \mu_1 v_1 = \sum_{l \in L} \alpha_l T(v_l) = \sum_{l \in L} \alpha_l \mu_l v_l.$$

This shows that

$$0 = \sum_{l \in L} \alpha_l (\mu_l - \mu_1) v_l.$$

Hence, $\alpha_l(\mu_l - \mu_1) = 0$ for all $l \in L$, which, in turn, implies that $\alpha_l = 0$ since $\mu_1 \neq \mu_l$ for all $l \in L$. In conclusion, $v_1 = 0$, which is a contradiction since v_1 is an eigenvector (i.e., v_1 cannot be equal to zero). Hence, the vectors $v_1, \dots, v_m$ are linearly independent.

Exercise 46.4 (Volterra operator). Let $L := L^2(D; \mathbb{C})$ with $D := (0, 1)$.
(i) Let $f, g \in L$. Integrating by parts, we obtain

$$\begin{aligned}(g, T(f))_L &= \int_0^1 g(x) \Big(\int_0^x \overline{f}(t) \, \mathrm{d}t \Big) \, \mathrm{d}x = \int_0^1 \partial_x \Big(\int_1^x g(t) \, \mathrm{d}t \Big) \Big(\int_0^x \overline{f}(t) \, \mathrm{d}t \Big) \, \mathrm{d}x \\ &= - \int_0^1 \Big(\int_1^x g(t) \, \mathrm{d}t \Big) \overline{f}(x) \, \mathrm{d}x.\end{aligned}$$

This means that $T^{\mathsf{H}}(g) = \int_x^1 g(t) \, \mathrm{d}t$. Note in passing the T is a Hilbert–Schmidt operator. Specifically, we have $T(f)(x) = \int_0^1 K(x, t) f(t) \, \mathrm{d}t$ with $K(x, t) := 1$ if $t \in (0, x)$ and $K(x, t) := 0$ otherwise (see Example 46.11). Hence, T is compact and not symmetric.
(ii) Let us show that T is injective. Assume that $T(f) = 0$. Then $\int_0^x f(t) \, \mathrm{d}t = 0$ for a.e. $x \in D$. To conclude that $f = 0$, we apply the vanishing integral theorem (Theorem 1.32) by showing that $(f, \varphi)_L = 0$ for all $\varphi \in C_0^\infty(D)$. Let $\varphi \in C_0^\infty(D)$ and let us define $\psi(x) := -\varphi'(x)$. Observe that $\psi \in L$ and $T^{\mathsf{H}}(\psi) = \int_x^1 \psi(t) \, \mathrm{d}t = \varphi(x) - \varphi(1) = \varphi(x)$ since φ is compactly supported in D. Then $0 = (T(f), \psi)_L = (f, T^{\mathsf{H}}(\psi))_L = (f, \varphi)_L$ for all $\varphi \in C_0^\infty(D)$.
(iii) Let $g \in L$ and assume that $T(f) = g$ with $f \in L$, i.e., we have $\int_0^x f(t) \, \mathrm{d}t = g(x)$ for a.e.

$x \in D$. This means that $f = \partial_x g$, i.e., f is the weak derivative of g. But this is not possible unless $g \in H^1(D;\mathbb{C})$ and $g(0) = 0$. In conclusion, T is not surjective. This proves that $0 \in \sigma_{\mathrm{c}}(T) \cup \sigma_{\mathrm{r}}(T)$. Note though that the above argument shows that $\mathrm{im}(T) = \{g \in H^1(D;\mathbb{C}) \mid g(0) = 0\}$, from which we conclude that $\overline{\mathrm{im}(T)} = L^2(D;\mathbb{C}) = L$. (Another way to prove $\overline{\mathrm{im}(T)} = L$ consists of proving that the only function $h \in L$ that satisfies $(h, T(f))_L = 0$ for all $f \in L$ is $h = 0$ and invoking Corollary C.15. We leave the details to the reader.) Hence, $0 \in \sigma_{\mathrm{c}}(T)$.
(iv) Assume $\sigma_{\mathrm{p}}(T) \neq \emptyset$, and let $\mu \in \sigma_{\mathrm{p}}(T)$ and $0 \neq f \in L$ s.t. $T(f) = \mu f$. We have $\mu \neq 0$ since T is injective. Moreover, we observe that

$$\begin{aligned}\Big(T(f) = \mu f\Big) &\iff \Big(- \mu^{-1} e^{-\mu^{-1}x} \int_0^x f(t)\,\mathrm{d}t + e^{-\mu^{-1}x} f(x) = 0\Big)\\ &\iff \Big(\partial_x\big(e^{-\mu^{-1}x} \int_0^x f(t)\,\mathrm{d}t\big) = 0\Big),\end{aligned}$$

which shows that $e^{-\mu^{-1}x} \int_0^x f(t)\,\mathrm{d}t$ should be constant, but this constant must be zero since $\lim_{x\downarrow 0} \int_0^x f(t)\,\mathrm{d}t = 0$. Hence, $\int_0^x f(t)\,\mathrm{d}t = 0$ for a.e. $x \in (0,1)$. We conclude that $f = 0$ by using as above the vanishing integral theorem. This is a contradiction. This proves that $\sigma_{\mathrm{p}}(T) = \emptyset$.
(v) Let $\mu \neq 0$. Since T is injective, we only need to prove that T is surjective. Let $g \in L$. Let us try to find $f \in L$ such that $T(f) - \mu f = g$. This is equivalent to

$$\begin{aligned}\Big(\int_0^x f(t)\,\mathrm{d}t - \mu f = g\Big) &\iff \Big(- \mu e^{-\mu x} \int_0^x f(t)\,\mathrm{d}t + e^{-\mu x} f(x) = -\mu e^{-\mu x} g(x)\Big)\\ &\iff \Big(\partial_x\Big(e^{-\mu x} \int_0^x f(t)\,\mathrm{d}t\Big) = -\mu e^{-\mu x} g(x)\Big)\\ &\iff \Big(e^{-\mu x} \int_0^x f(t)\,\mathrm{d}t = -\mu \int_0^x e^{-\mu t} g(t)\,\mathrm{d}t\Big)\\ &\iff \Big(f(x) = -\mu^2 e^{\mu x} \int_0^x e^{-\mu t} g(t)\,\mathrm{d}t - \mu g(x)\Big).\end{aligned}$$

The triangle inequality and the Cauchy–Schwarz inequality imply that $f \in L = L^2((0,1);\mathbb{C})$, and there is a constant c that depends on μ such that $\|f\|_L \leq c\|g\|_L$. This proves that $T - \mu I_L$ is bijective if $\mu \neq 0$. (Notice that T is compact since it is a Hilbert–Schmidt operator; see Example 46.11). Hence, we could also invoke Theorem 46.14 (i)-(ii) which implies that $\{0\} = \sigma_{\mathrm{c}}(T) \cup \sigma_{\mathrm{r}}(T)$. Since we have already shown that $\sigma_{\mathrm{p}}(T) = \emptyset$, we conclude that $\rho(T) = \mathbb{C}\backslash\{0\}$, i.e., $\mu I_L - T$ is bijective for all $\mu \neq 0$.)
(vi) We have shown that $\rho(T) = \mathbb{C}\backslash\{0\}$, $\sigma_{\mathrm{p}}(T) = \emptyset$, $\sigma_{\mathrm{c}}(T) = \{0\}$, $\sigma_{\mathrm{r}}(T) = \emptyset$. The same results hold true for T^{H}.

Exercise 46.5 (Riesz–Fréchet). (i) Let $g \in H'$. Let $y := \sum_{i\in\{1:n\}} y_i e_i \in H$ and let $x_g := \sum_{i\in\{1:n\}} g(e_i) e_i \in H$ with $g(e_i) := \langle g, e_i\rangle_{H',H}$. Using the orthonormality of the Hilbert basis, we obtain

$$\begin{aligned}(x_g, y)_H &= \sum_{i\in\{1:n\}} (g(e_i)e_i, y_i e_i)_H = \sum_{i\in\{1:n\}} g(e_i)\overline{y_i}\\ &= \sum_{i\in\{1:n\}} \langle g, y_i e_i\rangle_{H',H} = \langle g, y\rangle_{H',H} =: ((J^{\mathrm{RF}})^{-1}(g), y)_H.\end{aligned}$$

This proves that $x_g = (J^{\mathrm{RF}})^{-1}(g)$. The map $g \mapsto x_g$ is clearly linear, i.e., $(J^{\mathrm{RF}})^{-1}$ and J^{RF} are linear operators.

(ii) Similarly, we have

$$(x_g, y)_H = \sum_{i\in\{1:n\}} (\overline{g(e_i)}e_i, y_i e_i)_H = \sum_{i\in\{1:n\}} \overline{g(e_i)y_i} = \overline{g(y)} := \overline{\langle g, y\rangle_{H',H}}.$$

Finally, the map $g \mapsto x_g$ is clearly antilinear.

Exercise 46.6 (Symmetric operator). (i) Notice first that $(T(v), v)_L = (v, T(v))_L = \overline{(T(v), v)_L}$ for all $v \in L$. Hence, $(T(v), v)_L \in \mathbb{R}$ for all $v \in L$. Let $\mu \in \mathbb{C}$. We have

$$\begin{aligned} 2\mathrm{i}\Im(T(v) - \mu v, v)_L &= (T(v) - \mu v, v)_L - \overline{(T(v) - \mu v, v)_L} \\ &= (T(v), v)_L - \mu\|v\|_L^2 - \overline{(T(v), v)_L} + \overline{\mu}\|v\|_L^2 \\ &= -2\mathrm{i}\Im(\mu)\|v\|_L^2. \end{aligned}$$

This proves that $|\Im(\mu)|\|v\|_L^2 \le |((T - \mu I_L)(v), v)_L|$. Hence, if $\Im(\mu) \neq 0$, then $T - \mu I_L$ is coercive, that is, $\Im(\mu) \neq 0$ implies that $\mu \in \rho(T) = \mathbb{C}\backslash\sigma(T)$. In other words, $\mu \in \sigma(T) = \mathbb{C}\backslash\rho(T)$ implies that $\mu \in \mathbb{R}$.
(ii) Assume that $\sigma_{\mathrm{r}}(T) \neq \emptyset$. Let $\mu \in \sigma_{\mathrm{r}}(T)$. Then $T - \mu I_L$ is injective and $\mathrm{im}(T - \mu I_L)$ is not dense in L. Corollary C.15 implies that there is $0 \neq f \in L$ such that $(f, T(v) - \mu v)_L = 0$ for all $v \in L$. Since $\mu \in \mathbb{R}$, this means that $(T(f) - \mu f, v)_L = 0$ for all $v \in L$. This, in turn, implies that $(T - \mu I_L)(f) = 0$, i.e., $\mu \in \sigma_{\mathrm{p}}(T)$, which is impossible since $\sigma_{\mathrm{p}}(T) \cap \sigma_{\mathrm{r}}(T) = \emptyset$. Hence, $\sigma_{\mathrm{r}}(T) = \emptyset$. (iii) Let $\mu \in \sigma_{\mathrm{p}}(T)$. Let $x \in \ker(\mu I_L - T)$. Then $(\mu I_L - T) \circ (\mu I_L - T)(x) = 0$, i.e., $x \in \ker(\mu I_L - T)^2$. This shows that $\ker(\mu I_L - T) \subset \ker(\mu I_L - T)^2$. Let $x \in \ker(\mu I_L - T)^2$. This means that $(\mu I_L - T) \circ (\mu I_L - T)(x) = 0$, and

$$\begin{aligned} 0 &= (x, (\mu I_L - T) \circ (\mu I_L - T)(x))_L = ((\overline{\mu} I_L - T^{\mathsf{H}})(x), (\mu I_L - T)(x))_L \\ &= ((\mu I_L - T)(x), (\mu I_L - T)(x))_L = \|(\mu I_L - T)(x)\|_L^2, \end{aligned}$$

where we used that $\overline{\mu} = \mu$ and $T^{\mathsf{H}} = T$. The above equality implies that $(\mu I_L - T)(x) \in \ker(\mu I_L - T)$, and this shows that $\ker(\mu I_L - T)^2 = \ker(\mu I_L - T)$, thereby proving that the ascent of μ is equal to 1.

Exercise 46.7 ($H^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is not compact). (i) Let $\chi(x) := 1 + x$ if $-1 \le x \le 0$, $\chi(x) := 1 - x$ if $0 \le x \le 1$, and $\chi(x) := 0$ if $|x| \ge 1$. It is clear that $\chi \in L^2(\mathbb{R})$. Moreover, the weak derivative of χ is equal to 1 if $-1 \le x \le 0$, -1 if $0 \le x \le 1$, and 0 if $|x| \ge 1$. Hence, $\chi \in H^1(\mathbb{R})$.
(ii) Consider the sequence $v_n(x) := \chi(x - n)$ for all $n \in \mathbb{N}$. Let $\phi \in L^2(\mathbb{R})$. We have

$$\left|\int_{\mathbb{R}} \phi(x) v_n(x)\,\mathrm{d}x\right| = \left|\int_{n-1}^{n+1} \phi(x)\chi(x - n)\,\mathrm{d}x\right| \le \|\phi\|_{L^2(n-1,n+1)}\sqrt{\frac{2}{3}}.$$

But $\|\phi\|_{L^2(n-1,n+1)} \le \|\phi\|_{L^2(n-1,\infty)} \to 0$ as $n \to \infty$. Hence, $\int_{\mathbb{R}} \phi(x)v_n(x)\,\mathrm{d}x \to 0$ as $n \to \infty$, for all $\phi \in L^2(\mathbb{R})$. According to Definition C.28, this means that the sequence $(v_n)_{n\in\mathbb{N}}$ converges weakly to 0 in $L^2(\mathbb{R})$.
(iii) We argue by contradiction. Assume that the embedding $H^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact. Since $\|v_n\|_{L^2(\mathbb{R})} = \sqrt{\frac{2}{3}}$ and $|v_n|_{H^1(\mathbb{R})} = \sqrt{2}$, the sequence $(v_n)_{n\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R})$. Owing to Theorem C.23, we infer that there exists a subsequence $(v_{n_k})_{k\in\mathbb{N}}$ that converges to some $v \in L^2(\mathbb{R})$. Since strong convergence implies weak convergence and $(v_n)_{n\in\mathbb{N}}$ converges weakly to zero owing to Step (ii), we must have $v = 0$, but since $\|v_n\|_{L^2(\mathbb{R})} = \sqrt{\frac{2}{3}}$, we must have $\|v\|_{L^2(\mathbb{R})} = \sqrt{\frac{2}{3}} > 0$. This is a contradiction.

Exercise 46.8 ($B^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact). (i) Let $(u_n)_{n\in\mathbb{N}}$ be a sequence in the unit ball of $B^1(\mathbb{R})$. Let us set $J_0 := \mathbb{N}$. Let $k \in \mathbb{N}\backslash\{0\}$. Using that the embedding $H^1(-k,k) \hookrightarrow L^2(-k,k)$ is compact, let us extract a subset $J_1 \subset J_0$ such that $(u_{n|(-1,1)})_{n\in J_1}$ converges strongly to some function v_1 in $L^2((-1,1))$. By induction on $k \geq 1$, we extract from the sequence $(u_n)_{n\in J_k}$ a subsequence $(u_n)_{n\in J_{k+1}}$ such that $(u_{n|(-(k+1),k+1)})_{n\in J_{k+1}}$ converges to some v_{k+1} in $L^2(-(k+1),k+1)$. Note that by construction $v_{k+1|(-k,k)} = v_k$ since the sequence $(u_n)_{n\in J_{k+1}}$ is a subsequence of $(u_n)_{n\in J_k}$. For each $k \in \mathbb{N}\backslash\{0\}$, we define n_k to be the smallest integer in J_k such that $\|u_{n_k} - u_m\|_{L^2(-k,k)} \leq \frac{1}{k}$ for all $m \in J_k$ such that $m \geq n_k$. This is legitimate since $(u_{n|(-k,k)})_{n\in J_k}$ is a Cauchy sequence in $L^2(-k,k)$. Note that $n_{k+1} \in J_{k+1} \subset J_k$ and for all $m \geq n_{k+1}$ we have $\|u_{n_{k+1}} - u_m\|_{L^2(-k,k)} \leq \|u_{n_{k+1}} - u_m\|_{L^2(-k-1,k+1)} \leq \frac{1}{k+1} \leq \frac{1}{k}$. Hence, $n_k \leq n_{k+1}$. As a result, we have for all $k \leq l \in \mathbb{N}$,

$$\begin{aligned}\|u_{n_k} - u_{n_l}\|_{L^2(\mathbb{R})} &\leq \|u_{n_k} - u_{n_l}\|_{L^2(-k,k)} + \|u_{n_k}\|_{L^2(\mathbb{R}\backslash(-k,k))} + \|u_{n_l}\|_{L^2(\mathbb{R}\backslash(-k,k))}\\ &\leq \frac{1}{k} + \frac{1}{k} + \frac{1}{l} \leq \frac{3}{k},\end{aligned}$$

where we used that $\|v\|_{L^2(\mathbb{R}\backslash(-k,k))} \leq \frac{1}{k}\|v\|_{B^1(\mathbb{R})}$ for all $v \in B^1(\mathbb{R})$. It follows that $(u_{n_k})_{k\in\mathbb{N}}$ is a Cauchy sequence in $L^2(\mathbb{R})$. This proves the compactness of the embedding $B^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$.
(ii) The above proof shows that the embedding $B^1_\alpha(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact if $\alpha > 0$ since in this case $\|v\|_{L^2(\mathbb{R}\backslash(-k,k))} \leq \frac{1}{k^\alpha}\|v\|_{B^1_\alpha(\mathbb{R})}$.

Exercise 46.9 (Hausdorff–Toeplitz theorem). (i) Using the proposed definitions, we have

$$\begin{aligned}(T'(x_1),x_1)_L &= \frac{1}{\mu-\gamma}((T(x_1),x_1)_L - \gamma(x_1,x_1)_L) = \frac{1}{\mu-\gamma}(\gamma-\gamma) = 0,\\ (T'(x_2),x_2)_L &= \frac{1}{\mu-\gamma}((T(x_2),x_2)_L - \gamma(x_2,x_2)_L) = \frac{1}{\mu-\gamma}(\mu-\gamma) = 1.\end{aligned}$$

(ii) Let us compute $\Im\big(e^{\mathrm{i}\theta}(T'(x_1),x_2)_L + e^{-\mathrm{i}\theta}(T'(x_2),x_1)_L\big)$. We have

$$\begin{aligned}\Im\big(e^{\mathrm{i}\theta}(T'(x_1),x_2)_L + e^{-\mathrm{i}\theta}(T'(x_2),x_1)_L\big) = {}& \cos(\theta)\Im\big((T'(x_1),x_2)_L\big) + \sin(\theta)\Re\big((T'(x_1),x_2)_L\big)\\ &+ \cos(\theta)\Im\big((T'(x_2),x_1)_L\big) - \sin(\theta)\Re\big((T'(x_1),x_2)_L\big).\end{aligned}$$

The equation $\Im\big(e^{\mathrm{i}\theta}(T'(x_1),x_2)_L + e^{-\mathrm{i}\theta}(T'(x_2),x_1)_L\big) = 0$ is equivalent to

$$\cos(\theta)\Im\big((T'(x_1),x_2)_L + (T'(x_2),x_1)_L\big) + \sin(\theta)\Re\big((T'(x_1),x_2)_L - (T'(x_1),x_2)_L\big) = 0.$$

This problem amounts to finding a unit vector $(\cos(\theta),\sin(\theta))^\mathsf{T}$ that is orthogonal to the vector $\big(\Im\big((T'(x_1),x_2)_L + (T'(x_2),x_1)_L\big), \Re\big((T'(x_1),x_2)_L - (T'(x_1),x_2)_L\big)\big)^\mathsf{T}$. There are two angles θ satisfying this property.
(iii) Let us set $x_1' := e^{\mathrm{i}\theta}x_1$. We obtain

$$(T'(x_1'),x_1')_L = e^{\mathrm{i}\theta}(T'(x_1),x_1')_L = e^{\mathrm{i}\theta}e^{-\mathrm{i}\theta}(T'(x_1),x_1)_L = 0.$$

(iv) Let $\lambda \in [0,1]$, and let us consider the following problem: Find $\alpha,\beta \in \mathbb{R}$ s.t. $\|\alpha x_1' + \beta x_2\|_L = 1$ and $(T'(\alpha x_1' + \beta x_2), \alpha x_1' + \beta x_2)_L = \lambda$. We have

$$\begin{aligned}1 &= \|\alpha x_1' + \beta x_2\|_L^2 = \alpha^2\|x_1'\|_L^2 + \beta^2\|x_2\|_L^2 + 2\alpha\beta\Re\big((x_1',x_2)_L\big),\\ &= \alpha^2 + \beta^2 + 2\alpha\beta\Re\big((x_1',x_2)_L\big).\end{aligned}$$

The set of points $(\alpha,\beta)\in\mathbb{R}^2$ satisfying this equation is an ellipse intersecting the axes at $(1,0)$, $(0,1)$, $(-1,0)$, and $(0,-1)$ (note that we used the Cauchy–Schwarz inequality here). Moreover, we observe that

$$\begin{aligned}\lambda &= (T'(\alpha x_1'+\beta x_2),\alpha x_1'+\beta x_2)_L\\ &= \alpha^2(T'(x_1'),x_1')+\beta^2(T'(x_2),x_2)_L+\alpha\beta((T'(x_1'),x_2)_L+(T'(x_2),x_1')_L)\\ &= \beta^2+\alpha\beta\Re\big((T'(x_1'),x_2)_L+(T'(x_2),x_1')_L\big).\end{aligned}$$

(Notice that we used $\Im\big((T'(x_1'),x_2)_L+(T'(x_2),x_1')_L\big)=0$ here.) The set of points $(\alpha,\beta)\in\mathbb{R}^2$ satisfying this equation is an hyperbola intersecting the vertical axis at $\pm\sqrt{\lambda}$. Since $\lambda\in[0,1]$, we conclude that the system

$$\begin{aligned}&\alpha^2+\beta^2+2\alpha\beta\Re\big((x_1',x_2)_L\big)=1,\\ &\beta^2+\alpha\beta\Re\big((T'(x_1),x_2)_L+(T'(x_2),x_1)_L\big)=\lambda,\end{aligned}$$

has at least two solutions (four in general).

(v) Let us prove that $W(T)$ is convex. Let $\gamma,\mu\in W(T)\subset\mathbb{C}$ and let us prove that the segment connecting γ to μ is in $W(T)$. There is nothing to prove if $\gamma=\mu$. Let us assume now that $\gamma\neq\mu$. Let $x_1,x_2\in L$ be s.t. $\|x_1\|_L=\|x_2\|_L:=1$ and $(T(x_1),x_1)_L:=\gamma$, $(T(x_2),x_2)_L:=\mu$. Let $\lambda\in[0,1]$, and let x_1', α, and β be constructed as above. We obtain

$$\begin{aligned}(T(\alpha x_1'+\beta x_2),\alpha x_1'+\beta x_2)_L &= (T(\alpha x_1'+\beta x_2)-\gamma(\alpha x_1'+\beta x_2),\alpha x_1'+\beta x_2)_L\\ &\quad+\gamma(\alpha x_1'+\beta x_2,\alpha x_1'+\beta x_2)_L\\ &= (\mu-\gamma)(T'(\alpha x_1'+\beta x_2),\alpha x_1'+\beta x_2)_L+\gamma\\ &= (\mu-\gamma)\lambda+\gamma.\end{aligned}$$

This proves that $(\mu-\gamma)\lambda+\gamma\in W(T)$ for all $\lambda\in[0,1]$ because $\|\alpha x_1'+\beta x_2\|_L=1$. Hence, $W(T)$ is convex.

Chapter 47

Symmetric operators, conforming approximation

Exercises

Exercise 47.1 (Real eigenvalues). Consider the eigenvalue problem: Find $\psi \in H^1_0(D;\mathbb{C})\backslash\{0\}$ and $\lambda \in \mathbb{C}$ s.t. $\int_D (\nabla\psi\cdot\nabla\overline{w} + \psi\overline{w})\,\mathrm{d}x = \lambda\int_D \psi\overline{w}\,\mathrm{d}x$ for all $w \in H^1_0(D;\mathbb{C})$. Prove directly that λ is real. (*Hint*: test with $w := \psi$.)

Exercise 47.2 (Smallest eigenvalue). Let $D_1 \subset D_2$ be two Lipschitz domains in $\mathbb{R}^d$. Let $a_i : H^1_0(D_i)\times H^1_0(D_i) \to \mathbb{R}$, $i \in \{1,2\}$, be two symmetric, coercive, bounded bilinear forms. Assume that $a_1(v,w) = a_2(\widetilde{v},\widetilde{w})$ for all $v, w \in H^1_0(D_1)$, where $\widetilde{v},\widetilde{w}$ denote the extension by zero of v, w, respectively. Let $\lambda_1(D_i)$ be the smallest eigenvalue of the eigenvalue problem: Find $\psi \in H^1_0(D_i)\backslash\{0\}$ and $\lambda \in \mathbb{R}$ s.t. $a_i(\psi,w) = \lambda(\psi,w)_{L^2(D_i)}$ for all $w \in H^1_0(D_i)$. Prove that $\lambda_1(D_2) \le \lambda_1(D_1)$. (*Hint*: use Proposition 47.3.)

Exercise 47.3 (Continuity of eigenvalues). Consider the setting defined in §47.1. Let $a_1, a_2 : V\times V \to \mathbb{R}$ be two symmetric, coercive, bounded bilinear forms. Let $A_1, A_2 : V \to V'$ be the linear operators defined by $\langle A_i(v), w\rangle_{V',V} := a_i(v,w)$, $i \in \{1,2\}$, for all $v, w \in A$. Let $\lambda_k(a_1)$ and $\lambda_k(a_2)$ be the k-th eigenvalues, respectively. Prove that $|\lambda_k(a_1) - \lambda_k(a_2)| \le \sup_{v\in S} |\langle (A_1 - A_2)(v), v\rangle_{V',V}|$, where S is the unit sphere in $L^2(D)$. (*Hint*: use the min-max principle.)

Exercise 47.4 (Max-min principle). Prove the second equality in (47.6). (*Hint*: let $E_{m-1} \in V_{m-1}$ and observe that $E^\perp_{m-1} \cap W_m \neq \{0\}$.)

Exercise 47.5 (Laplacian, 1D). Consider the spectral problem for the 1D Laplacian on $D := (0,1)$. (i) Show that the eigenpairs (λ_m, ψ_m) are $\lambda_m = m^2\pi^2$, $\psi_m(x) = \sin(m\pi x)$, for all $x \in D$ and all $m \ge 1$. (ii) Consider a uniform mesh of D of size $h := \frac{1}{I+1}$ and H^1-conforming $\mathbb{P}_1$ finite elements. Compute the stiffness matrix $\mathcal{A}$ and the mass matrix $\mathcal{M}$. (iii) Show that the eigenvalues of the discrete problem (47.8) are $\lambda_{hm} = \frac{6}{h^2}(\frac{1-\cos(m\pi h)}{2+\cos(m\pi h)})$ for all $m \in \{1{:}I\}$. (*Hint*: consider the vectors $(\sin(\pi hml))_{l\in\{1:I\}}$ for all $m \in \{1{:}I\}$.)

Exercise 47.6 (Stiffness matrix). Assume that the mesh sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is quasi-uniform. Estimate from below the smallest eigenvalue of the stiffness matrix $\mathcal{A}$ defined in (47.9) and estimate from above its largest eigenvalue. (*Hint*: see §28.2.3.)

Solution to exercises

Exercise 47.1 (Real eigenvalues). Testing with $w := \psi$ yields

$$\int_D (\|\nabla\psi\|_{\ell^2}^2 + |\psi|^2)\,\mathrm{d}x = \lambda \int_D |\psi|^2\,\mathrm{d}x.$$

Since $\psi \neq 0$, this shows that λ is real.

Exercise 47.2 (Smallest eigenvalue). For all $v \in H_0^1(D_1)$, let us denote by $\widetilde{v}$ the extension by zero of v over D_2. Theorem 3.18 implies that $\widetilde{v} \in H_0^1(D_2)$ and $\|\widetilde{v}\|_{H^1(D_2)} = \|v\|_{H^1(D_1)}$. We conclude using Proposition 47.3, which implies that

$$\begin{aligned}\lambda_1(D_1) &= \min_{v\in H_0^1(D_1)} \frac{a_1(v,v)}{\|v\|_{L^2(D_1)}^2} = \min_{v\in H_0^1(D_1)} \frac{a_2(\widetilde{v},\widetilde{v})}{\|\widetilde{v}\|_{L^2(D_2)}^2} \\ &\geq \min_{v\in H_0^1(D_2)} \frac{a_2(v,v)}{\|v\|_{L^2(D_2)}^2} = \lambda_1(D_2).\end{aligned}$$

Exercise 47.3 (Continuity of eigenvalues). Using the min-max principle (Proposition 47.4), we infer that

$$\lambda_k(a_1) = \min_{E_k\in V_k} \max_{v\in E_k} \frac{\langle A_1(v),v\rangle_{V',V}}{\|v\|_{L^2(D)}^2}, \qquad \lambda_k(a_2) = \min_{E_k\in V_k} \max_{v\in E_k} \frac{\langle A_2(v),v\rangle_{V',V}}{\|v\|_{L^2(D)}^2}.$$

Let $E_k^2 := \mathrm{span}\{\psi_1^2,\dots,\psi_k^2\}$, where $\{\psi_1^2,\dots,\psi_k^2\}$ are the k first eigenfunctions of a_2, so that $\lambda_k(a_2) = \max_{v\in E_k^2} \frac{\langle A_2(v),v\rangle_{V',V}}{\|v\|_{L^2(D)}^2}$. We obtain

$$\lambda_k(a_1) - \lambda_k(a_2) \leq \max_{v\in E_k^2} \frac{\langle A_1(v),v\rangle_{V',V}}{\|v\|_{L^2(D)}^2} - \max_{v\in E_k^2} \frac{\langle A_2(v),v\rangle_{V',V}}{\|v\|_{L^2(D)}^2}.$$

Let $g \in E_k^2 \setminus \{0\}$ be such that $\frac{\langle A_1(g),g\rangle_{V',V}}{\|g\|_{L^2(D)}^2} := \max_{v\in E_k^2} \frac{\langle A_1(v),v\rangle_{V',V}}{\|v\|_{L^2(D)}^2}$. We infer that

$$\lambda_k(a_1) - \lambda_k(a_2) \leq \frac{\langle A_1(g),g\rangle_{V',V}}{\|g\|_{L^2(D)}^2} - \frac{\langle A_2(g),g\rangle_{V',V}}{\|g\|_{L^2(D)}^2} = \frac{\langle (A_1-A_2)(g),g\rangle_{V',V}}{\|g\|_{L^2(D)}^2}.$$

Hence, $\lambda_k(a_1) - \lambda_k(a_2) \leq \sup_{v\in V} \frac{|\langle (A_1-A_2)(v),v\rangle_{V',V}|}{\|v\|_{L^2(D)}^2}$. The other inequality is shown by switching the roles of $\lambda_k(a_1)$ and $\lambda_k(a_2)$.

Exercise 47.4 (Max-min principle). Let us set

$$W_{m-1} := \mathrm{span}\{\psi_1,\dots,\psi_{m-1}\}, \qquad W_m := \mathrm{span}\{\psi_1,\dots,\psi_m\}.$$

For all $v \in W_{m-1}^\perp$, we have $v \in \mathrm{span}\{\psi_n\}_{n\geq m}$, so that

$$\max_{E_{m-1}\in V_{m-1}} \min_{v\in E_{m-1}^\perp} R(v) \geq \min_{v\in W_{m-1}^\perp} R(v) = \min_{v\in W_{m-1}^\perp} \frac{\sum_{n\geq m}\lambda_n v_n^2}{\sum_{n\geq m} v_n^2} = \lambda_m.$$

Let now $E_{m-1} \in V_{m-1}$. A dimension argument shows that $E_{m-1}^\perp \cap W_m \neq \{0\}$. Thus, we have

$$\min_{v\in E_{m-1}^\perp} R(v) \leq \min_{v\in E_{m-1}^\perp\cap W_m} R(v) \leq \max_{v\in E_{m-1}^\perp\cap W_m} R(v) \leq \max_{v\in W_m} R(v) = \lambda_m,$$

where the last equality follows from the min-max principle. This proves that

$$\max_{E_{m-1}\in V_{m-1}} \min_{v\in E_{m-1}^{\perp}} R(v) \le \lambda_m.$$

Altogether, we have shown that

$$\max_{E_{m-1}\in V_{m-1}} \min_{v\in E_{m-1}^{\perp}} R(v) = \lambda_m.$$

Exercise 47.5 (Laplacian, 1D). (i) The spectral problem for the Laplacian in the domain $D := (0,1)$ is:

$$\begin{cases} \text{Find } \psi \in H_0^1(D)\backslash\{0\} \text{ and } \lambda \in \mathbb{R} \text{ such that} \\ a(\psi, w) = \lambda(\psi, w)_{L^2(D)}, \quad \forall w \in H_0^1(D), \end{cases}$$

where $a(\psi, w) := \int_0^1 \psi' w' dx$. It follows that $-\psi'' = \lambda\psi$, $\psi \in H_0^1(D)$. This is an ordinary differential equation with characteristic equation $-s^2 = \lambda$. If $\lambda < 0$, we have $s = \pm\sqrt{-\lambda}$, and the fundamental solutions are $e^{\sqrt{-\lambda}x}$ and $e^{-\sqrt{-\lambda}x}$. But these two fundamental solutions do not satisfy the boundary condition. Hence, $\lambda \ge 0$, and the two fundamental solutions are $\cos(\sqrt{\lambda}x)$ and $\sin(\sqrt{\lambda}x)$. It follows from $\psi(0) = \psi(1) = 0$ that $\psi(x) = \sin(m\pi x)$ with $\lambda = m^2\pi^2$ and $m \in \mathbb{N}\backslash\{0\}$ (since $\psi \neq 0$).
(ii) The discrete eigenvalue problem is

$$\begin{cases} \text{Find } \psi_h \in V_h\backslash\{0\} \text{ and } \lambda_h \in \mathbb{R} \text{ such that} \\ a(\psi_h, w_h) = \lambda_h(\psi_h, w_h)_{L^2(D)}, \quad \forall w_h \in V_h, \end{cases}$$

where $V_h := P^{\mathrm{g}}_{1,0}(\mathcal{T}_h)$. This problem can be recast as follows:

$$\begin{cases} \text{Find } \mathsf{U}_h \in \mathbb{R}^I\backslash\{0\} \text{ and } \lambda_h \in \mathbb{R} \text{ such that} \\ \mathcal{A}\mathsf{U}_h = \lambda_h \mathcal{M}\mathsf{U}_h, \end{cases}$$

where $\mathcal{A}_{ij} := a(\varphi_j, \varphi_i)$, $\mathcal{M}_{ij} := (\varphi_j, \varphi_i)_{L^2(D)}$ and $\{\varphi_1, \cdots, \varphi_I\}$ are the global shape functions in V_h. For all $i \in \{1{:}I\}$, we have

$$\varphi_i(x) := \begin{cases} -\frac{1}{h}|x - x_i| + 1 & \text{if } x \in [x_{i-1}, x_{i+1}], \\ 0 & \text{otherwise.} \end{cases}$$

We infer that

$$\mathcal{A}_{ij} = \int_0^1 \varphi_j'(x)\varphi_i'(x)dx = \begin{cases} \frac{2}{h} & \text{if } i = j, \\ -\frac{1}{h} & \text{if } |i-j| = 1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mathcal{M}_{ij} = \int_0^1 \varphi_j(x)\varphi_i(x)dx = \begin{cases} \frac{4h}{6} & \text{if } i = j, \\ \frac{h}{6} & \text{if } |i-j| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $\mathcal{A} = \frac{1}{h}\mathrm{tridiag}(-1, 2, -1)$ and $\mathcal{M} = \frac{h}{6}\mathrm{tridiag}(1, 4, 1)$.
(iii) Using that

$$\Im(2e^{\mathrm{i}\pi hml} - e^{\mathrm{i}\pi hm(l-1)} - e^{\mathrm{i}\pi hm(l+1)}) = 2(1 - \cos(\pi hm))\Im(2e^{\mathrm{i}\pi hml}),$$

we infer that the eigenvalues of the stiffness matrix $\mathcal{A}$ are

$$\lambda_{hm,\mathcal{A}} = \frac{2}{h}(1 - \cos(\pi hm)),$$

with the corresponding eigenvectors $\mathsf{U}_{hm} = (\sin(\pi hml))_{l\in\{1:I\}}^{\mathsf{T}}$ for all $m \in \{1:I\}$. Using that

$$\Im(4e^{\mathrm{i}\pi hml} + e^{\mathrm{i}\pi hm(l-1)} + e^{\mathrm{i}\pi hm(l+1)}) = (4 + 2\cos(\pi hm))\Im(2e^{\mathrm{i}\pi hml}),$$

we infer that the eigenvalues of the mass matrix $\mathcal{M}$ are

$$\lambda_{hm,\mathcal{M}} = \frac{h}{3}(2 + \cos(\pi hm)),$$

with the corresponding eigenvectors $\mathsf{U}_{hm} = (\sin(\pi hml))_{l\in\{1:I\}}^{\mathsf{T}}$ for all $m \in \{1:I\}$. The identity

$$\lambda_{hm,\mathcal{A}}\mathsf{U}_{hm} = \mathcal{A}\mathsf{U}_{hm} = \lambda_{hm}\mathcal{M}\mathsf{U}_{hm} = \lambda_{hm}\lambda_{hm,\mathcal{M}}\mathsf{U}_{hm},$$

for all $m \in \{1:I\}$ shows that $\lambda_{hm} = \frac{\lambda_{hm,\mathcal{A}}}{\lambda_{hm,\mathcal{M}}} = \frac{6}{h^2}(\frac{1-\cos(m\pi h)}{2+\cos(m\pi h)})$.

Exercise 47.6 (Stiffness matrix). We are going to use Proposition 28.11. We first have

$$\alpha_{L^2} := \inf_{v_h\in V_h}\sup_{w_h\in W_h}\frac{|a(v_h,w_h)|}{\|v_h\|_{L^2(D)}\|w_h\|_{L^2(D)}} \geq \inf_{v_h\in V_h}\frac{|a(v_h,v_h)|}{\|v_h\|_{L^2(D)}^2} \geq \alpha.$$

We also have

$$\begin{aligned}\omega_{L^2} &:= \sup_{v_h\in V_h}\sup_{w_h\in W_h}\frac{|a(v_h,w_h)|}{\|v_h\|_{L^2(D)}\|w_h\|_{L^2(D)}}\\ &\leq \|a\|\left(\sup_{v_h\in V_h}\frac{\|v_h\|_{H^1(D)}}{\|v_h\|_{L^2(D)}}\right)^2 \leq c\,\|a\|\ell_D^2 h^{-2}.\end{aligned}$$

Let $\mu_{\mathcal{M}}^{\min}$, $\mu_{\mathcal{M}}^{\max}$ be the smallest and the largest eigenvalues of the mass matrix $\mathcal{M}$, respectively. Owing to Proposition 28.6, we infer that $c_1h^d \leq \mu_{\mathcal{M}}^{\min} \leq \mu_{\mathcal{M}}^{\max} \leq c_2h^d$. Let $\lambda_{\mathcal{A}}^{\min}$, $\lambda_{\mathcal{A}}^{\max}$ be the smallest and the largest eigenvalues of $\mathcal{A}$, respectively. Since $\mathcal{A}$ is symmetric, we have $\lambda_{\mathcal{A}}^{\min} = \|\mathcal{A}^{-1}\|_{\ell^2}^{-1}$ and $\lambda_{\mathcal{A}}^{\max} = \|\mathcal{A}\|_{\ell^2}$. Finally, Proposition 28.11 gives

$$\lambda_{\mathcal{A}}^{\min} = \|\mathcal{A}^{-1}\|_{\ell^2}^{-1} \geq \mu_{\mathcal{M}}^{\min}\alpha_{L^2} \geq \alpha\mu_{\mathcal{M}}^{\min} \geq c\,\alpha h^d,$$

and

$$\lambda_{\mathcal{A}}^{\max} = \|\mathcal{A}\|_{\ell^2} \leq \mu_{\mathcal{M}}^{\max}\omega_{L^2} \leq c\,\mu_{\mathcal{M}}^{\max}\|a\|\ell_D^2h^{-2} \leq c\,\|a\|\ell_D^2h^{d-2}.$$

Chapter 48

Nonsymmetric problems

Exercises

Exercise 48.1 (Linearity). Consider the setting of §48.1.2. Let $V \hookrightarrow L$ be two complex Banach spaces and $a : V \times V \to \mathbb{C}$ be a bounded sesquilinear form satisfying the two conditions of the BNB theorem. Let $b : L \times L \to \mathbb{C}$ be bounded sesquilinear form. (i) Let $T : L \to L$ be such that $a(T(v), w) := b(v, w)$ for all $v \in L$ and all $w \in V$. Show that T is well defined and linear. (ii) Let $T_* : L \to L$ be such that $a(v, T_*(w)) := b(v, w)$ for all $v \in V$ and all $w \in L$. Show that T_* is well defined and linear.

Exercise 48.2 (Invariant sets). (i) Let $S, T \in \mathcal{L}(V)$ be such that $ST = TS$. Prove that $\ker(S)$ and $\mathrm{im}(S)$ are invariant under T. (ii) Let $T \in \mathcal{L}(V)$ and let $W_1, \ldots, W_m$ be subspaces of V that are invariant under T. Prove that $W_1 + \ldots + W_m$ and $\bigcap_{i \in \{1:m\}} W_i$ are invariant under T. (iii) Let $T \in \mathcal{L}(V)$ and let $\{v_1, \ldots, v_n\}$ be a basis of V. Show that the following statements are equivalent: (a) The matrix of T with respect to $\{v_1, \ldots, v_n\}$ is upper triangular; (b) $T(v_j) \in \mathrm{span}\{v_1, \ldots, v_j\}$ for all $j \in \{1{:}n\}$; (c) $\mathrm{span}\{v_1, \ldots, v_j\}$ is invariant under T for all $j \in \{1{:}n\}$. (iv) Let $T \in \mathcal{L}(V)$. Let μ be an eigenvalue of T. Prove that $\mathrm{im}(\mu I_V - T)$ is invariant under T. Prove that $\ker(\mu I_V - T)^\alpha$ is invariant under T for every integer $\alpha \geq 1$.

Exercise 48.3 (Trace). (i) Let V be a complex Banach space. Let $G \subset V$ be a subspace of V of dimension m. Let $\{\phi_j\}_{j \in \{1:m\}}$ and $\{\psi_j\}_{j \in \{1:m\}}$ be two bases of G, and let $\{\phi'_j\}_{j \in \{1:m\}}$ and $\{\psi'_j\}_{j \in \{1:m\}}$ be corresponding dual bases, i.e., $\langle \phi'_i, \phi_j \rangle_{V',V} = \delta_{ij}$, etc. (the way the antilinear forms $\{\phi'_j\}_{j \in \{1:m\}}$ and $\{\psi'_j\}_{j \in \{1:m\}}$ are extended to V does not matter). Let $T \in \mathcal{L}(V)$ and assume that G is invariant under T. Show that $\sum_{j \in \{1:m\}} \langle \psi'_j, T(\psi_j) \rangle_{V',V} = \sum_{j \in \{1:m\}} \langle \phi'_j, T(\phi_j) \rangle_{V',V}$. (ii) Let $B \in \mathbb{C}^{m \times m}$ be s.t. $T(\phi_i) =: \sum_{j \in \{1:m\}} B_{ji} \phi_j$ (recall that G is invariant under T). Let $\mathsf{V} := (\langle \phi'_j, v \rangle_{V',V})_{j \in \{1:m\}}^{\mathsf{T}}$ for all $v \in G$. Prove that $T^\alpha(v) = \sum_{j \in \{1:m\}} (B^\alpha \mathsf{V})_j \phi_j$ for all $\alpha \in \mathbb{N}$. (*Hint*: use an induction argument.) (iii) Let $\mu \in \mathbb{C}$, $\alpha \geq 1$, and $S \in \mathcal{L}(V)$. Assume that $G := \ker(\mu I_V - S)^\alpha$ is finite-dimensional and nontrivial (i.e., $\dim(G) := m \geq 1$). Prove that $\sum_{j \in \{1:m\}} \langle \phi'_j, S(\phi_j) \rangle_{V,V'} = m\mu$. (*Hint*: consider the $m \times m$ matrix A with entries $\langle \phi'_i, (\mu I_V - S)(\phi_j) \rangle_{V',V}$ and show that $A^\alpha = 0$.)

Exercise 48.4 (Theorem 48.12). Prove the estimates in Theorem 48.12. (*Hint*: see the proof of Theorem 48.8.)

Exercise 48.5 (Nonconforming approximation). Consider the Laplace operator with homogeneous Dirichlet boundary conditions in a Lipschitz polyhedron D with $b(v,w) := \int_D \rho v w \,\mathrm{d}x$, where $\rho \in C^\infty(D;\mathbb{R})$. Verify that the assumptions (48.25) to (48.30) hold true for the Crouzeix–Raviart approximation.

Solution to exercises

Exercise 48.1 (Linearity). (i) Let $\iota_{L,V} := \sup_{w\in V} \frac{\|w\|_L}{\|w\|_V}$. We first observe that $|b(v,w)| \le \|b\|\,\|v\|_L\|w\|_L \le \iota_{L,V}\|b\|\,\|v\|_L\|w\|_V$, that is, the antilinear form $f_v : V \to \mathbb{C}$ defined by $f_v(w) := b(v,w)$ is bounded. Then, for all $v \in L$, there exists a unique $T(v) \in V \hookrightarrow L$ s.t. $a(T(v),w) := f_v(w)$ for all $w \in V$. Let $v_1, v_2 \in V$ and $\alpha_1, \alpha_2 \in \mathbb{C}$. We obtain

$$\begin{aligned} a(T(\alpha_1 v_1 + \alpha_2 v_2), w) &= b(\alpha_1 v_1 + \alpha_2 v_2, w) = \alpha_1 b(v_1,w) + \alpha_2 b(v_2,w) \\ &= \alpha_1 a(T(v_1),w) + \alpha_2 a(T(v_2),w) \\ &= a(\alpha_1 T(v_1) + \alpha_2 T(v_2), w), \quad \forall w \in V. \end{aligned}$$

This means that $T(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 T(v_1) + \alpha_2 T(v_2)$, i.e., $T : L \to L$ is linear.
(ii) Using the same arguments as above, we prove that the linear form $g_w : V \to \mathbb{C}$ defined by $g_w(v) := b(v,w)$ is continuous. Then, for all $w \in L$, there exists a unique $T_*(w) \in V \hookrightarrow L$ s.t. $a(v, T_*(w)) := g_w(v)$ for all $v \in V$. Let $w_1, w_2 \in V$ and $\alpha_1, \alpha_2 \in \mathbb{C}$. We have

$$\begin{aligned} a(v, T(\alpha_1 w_1 + \alpha_2 w_2)) &= b(v, \alpha_1 w_1 + \alpha_2 w_2) = \overline{\alpha}_1 b(v,w_1) + \overline{\alpha}_2 b(v,w_2) \\ &= \overline{\alpha}_1 a(v, T_*(w_1)) + \overline{\alpha}_2 a(v, T_*(w_2)) \\ &= a(v, \alpha_1 T(w_1) + \alpha_2 T(w_2)), \quad \forall v \in V. \end{aligned}$$

This means that $T_*(\alpha_1 w_1 + \alpha_2 w_2) = \alpha_1 T_*(w_1) + \alpha_2 T_*(w_2)$, i.e., $T_* : L \to L$ is linear.

Exercise 48.2 (Invariant sets). (i) Let $S, T \in \mathcal{L}(V)$ be such that $ST = TS$. Let $v \in \ker(S)$ so that $ST(v) = TS(v) = 0$. Hence, $T(v) \in \ker(S)$, i.e., $\ker(S)$ is invariant under T. Let $v \in \mathrm{im}(S)$, i.e., there is $z \in V$ such that $v = S(z)$. This implies that $T(v) = TS(z) = ST(z) \in \mathrm{im}(S)$, i.e., $\mathrm{im}(S)$ is invariant under T.
(ii) Let $T \in \mathcal{L}(V)$ and let $W_1, \ldots, W_m$ be subspaces of V that are invariant under T. Let $w_1 + \ldots + w_m \in W_1 + \ldots + W_m$. Then $T(w_1 + \ldots + w_m) = T(w_1) + \ldots + T(w_m) \in W_1 + \ldots + W_m$, i.e., $W_1 + \ldots + W_m$ is invariant under T. Let $w \in \bigcap_{i\in\{1:m\}} W_i$. Then $T(w) \in W_i$, since $w \in W_i$ and W_i is invariant under T for all $i \in \{1{:}m\}$. Hence, $T(v) \in \bigcap_{i\in\{1:m\}} W_i$, i.e., $\bigcap_{i\in\{1:m\}} W_i$ is invariant under T.
(iii) We only prove that (b) implies (c) since the other implications are evident. Let us assume that (b) holds true. Let us fix $j \in \{1{:}n\}$. The statement (b) implies that $T(v_1) \in \mathrm{span}\{v_1\} \subset \mathrm{span}\{v_1, \ldots, v_n\}$, $T(v_2) \in \mathrm{span}\{v_1, v_2\} \subset \mathrm{span}\{v_1, \ldots, v_n\}$, $\ldots$, $T(v_n) \in \mathrm{span}\{v_1, \ldots, v_n\}$. Hence, if v is a linear combination of $v_1, \ldots, v_n$, then $T(v) \in \mathrm{span}\{v_1, \ldots, v_n\}$. In conclusion, we have shown that $\mathrm{span}\{v_1, \ldots, v_n\}$ is invariant under T, thereby proving (c).
(iv) Let μ be an eigenvalue of T. Let $v \in \mathrm{im}(\mu I_V - T)$. Then

$$T(v) = (T - \mu I_V)(v) + \mu v \in \mathrm{im}(\mu I_V - T) + \mathrm{span}\{v\} \subset \mathrm{im}(\mu I_V - T),$$

i.e., $\mathrm{im}(\mu I_V - T)$ is invariant under T. Let now $v \in \ker(\mu I_V - T)^\alpha$. We have

$$(\mu I_V - T)^\alpha (\mu I_V - T)(v) = (\mu I_V - T)(\mu I_V - T)^\alpha (v) = 0.$$

Hence, $\mu v - T(v) \in \ker(\mu I_V - T)^\alpha$. This implies that

$$T(v) \in \ker(\mu I_V - T)^\alpha + \operatorname{span}\{v\} = \ker(\mu I_V - T)^\alpha.$$

Hence, $\ker(\mu I_V - T)^\alpha$ is invariant under T.

Exercise 48.3 (Trace). (i) Since $\{\phi_j\}_{j\in\{1:m\}}$ and $\{\psi_j\}_{j\in\{1:m\}}$ are two bases of the same vector space, there exists an invertible $m\times m$ matrix A such that $\phi_j = \sum_{k\in\{1:m\}} A_{jk}\psi_k$ for all $j \in \{1{:}m\}$. Let $(A^{-1}_{ij})_{i,j\in\{1:m\}}$ be the coefficients of A^{-1} and let $\{\psi'_j\}_{j\in\{1:m\}}$ be a dual basis of $\{\psi_j\}_{j\in\{1:m\}}$. We obtain

$$\begin{aligned}\langle \sum_{k'\in\{1:m\}} \overline{A^{-1}_{k'i}}\psi'_{k'}, \phi_j\rangle_{V',V} &= \langle \sum_{k'\in\{1:m\}} \overline{A^{-1}_{k'i}}\psi'_{k'}, \sum_{k\in\{1:m\}} A_{jk}\psi_k\rangle_{V',V}\\ &= \sum_{k'\in\{1:m\}}\sum_{k\in\{1:m\}} \overline{A^{-1}_{k'i}}\overline{A_{jk}}\langle\psi'_{k'},\psi_k\rangle_{V',V}\\ &= \sum_{k\in\{1:m\}} \overline{A_{jk}}\overline{A^{-1}_{ki}} = \delta_{ji}.\end{aligned}$$

This proves that $\phi'_{i|G} = \sum_{k'\in\{1:m\}} \overline{A^{-1}_{k'i}}\psi'_{k'|G}$. Using that $T(G)\subset G$, i.e., that G is invariant under T, we infer that

$$\begin{aligned}\sum_{i\in\{1:m\}}\langle\phi'_i, T(\phi_i)\rangle_{V',V} &= \sum_{i\in\{1:m\}}\langle\phi'_{i|G}, T(\phi_i)\rangle_{V',V}\\ &= \sum_{i\in\{1:m\}}\langle \sum_{k'\in\{1:m\}} \overline{A^{-1}_{k'i}}\psi'_{k'|G}, \sum_{k\in\{1:m\}} A_{ik}T(\psi_k)\rangle_{V',V}\\ &= \sum_{k'\in\{1:m\}}\sum_{k\in\{1:m\}}\langle\psi'_{k'}, T(\psi_k)\rangle_{V',V}\sum_{i\in\{1:m\}}\overline{A^{-1}_{k'i}}\overline{A_{ik}}\\ &= \sum_{k\in\{1:m\}}\langle\psi'_k, T(\psi_k)\rangle_{V',V},\end{aligned}$$

which proves the expected result.
(ii) Since G is invariant under T, there are m^2 scalars $(B_{ij})_{i,j\in\{1:m\}}$ such that

$$T(\phi_i) = \sum_{j\in\{1:m\}} B_{ji}\phi_j.$$

Using the properties of the dual basis, we obtain

$$\langle\phi'_k, T(\phi_i)\rangle_{V',V} = \sum_{j\in\{1:m\}} B_{ji}\langle\phi'_k,\phi_j\rangle_{V',V} = B_{ki}.$$

Let $v := \sum_{i\in\{1:m\}} \mathsf{V}_i\phi_i \in G$. We have

$$T(v) = \sum_{i\in\{1:m\}} \mathsf{V}_i T(\phi_i) = \sum_{k\in\{1:m\}}\sum_{i\in\{1:m\}} B_{ki}\mathsf{V}_i\phi_k = \sum_{k\in\{1:m\}} (B\mathsf{V})_k\phi_k.$$

We can now conclude by using an induction argument as follows:

$$\begin{aligned} T^\alpha(v) &= \sum_{i\in\{1:m\}} (B^{\alpha-1}\mathsf{V})_i T(\phi_i) \\ &= \sum_{k\in\{1:m\}} \sum_{i\in\{1:m\}} B_{ki}(B^{\alpha-1}\mathsf{V})_i \phi_k \\ &= \sum_{k\in\{1:m\}} (B^\alpha \mathsf{V})_k \phi_k. \end{aligned}$$

(iii) Let us set $T := \mu I_V - S$. Let $v \in G := \ker(\mu I_V - S)^\alpha$, so that

$$\begin{aligned} (\mu I_V - S)^\alpha T(v) &= (\mu I_V - S)^\alpha (\mu I_V - S)(v) \\ &= (\mu I_V - S)(\mu I_V - S)^\alpha (v) = 0, \end{aligned}$$

which means that G is invariant under T. Let A be the $m\times m$ matrix with entries $\langle \phi_i', (\mu I_V - S)(\phi_j)\rangle_{V',V}$ for all $i,j \in \{1{:}m\}$. Since $T^\alpha(v) = (\mu I_V - S)^\alpha(v) = 0$ for all $v \in G$, the argument in Step (ii) shows that $A^\alpha \mathsf{V} = 0$ for all $\mathsf{V} \in \mathbb{C}^m$, i.e., $A^\alpha = 0$. Hence, the matrix A is nilpotent. Since the trace of any nilpotent matrix is zero, we infer that $\mathrm{tr}(A) = m\mu - \sum_{i\in\{1:m\}} \langle \phi_i', S(\phi_i)\rangle_{V',V} = 0$. We have thus proved that

$$\mathrm{tr}(S) := \sum_{i\in\{1:m\}} \langle \phi_i', S(\phi_i)\rangle_{V',V} = m\mu,$$

which is the expected result.

Exercise 48.4 (Theorem 48.12). We proceed as in the proof of Theorem 48.8. Using $t = \tau_\mu$ and $t^* = \tau^*$ in (48.31), we infer that

$$\|(T - T_h)_{|G_\mu}\|_{\mathcal{L}(G_\mu;L^2)} = \sup_{v\in G_\mu} \sup_{w\in L^2} \frac{((T-T_h)(v), w)_{L^2(D)}}{\|v\|_{L^2}\|w\|_{L^2}} \le c\, h^{\tau_\mu+\tau^*}.$$

Using $t = \tau$ and $t^* = \tau_\mu^*$ in (48.31), and recalling that $T^* = T^{\mathsf{H}}$, we infer that

$$\begin{aligned} \|(T - T_h)^*|_{G_\mu^*}\|_{\mathcal{L}(G_\mu^*;L^2)} &= \sup_{v\in L^2} \sup_{w\in G_\mu^*} \frac{(v, (T^{\mathsf{H}} - T_h^{\mathsf{H}})(w))_{L^2(D)}}{\|v\|_{L^2}\|w\|_{L^2}} \\ &= \sup_{v\in L^2} \sup_{w\in G_\mu^*} \frac{((T-T_h)(v), w)_{L^2(D)}}{\|v\|_{L^2}\|w\|_{L^2}} \le c\, h^{\tau+\tau_\mu^*}. \end{aligned}$$

Using $t = \tau_\mu$ and $t^* = \tau_\mu^*$ in (48.31), we finally infer that

$$\sup_{v\in G_\mu} \sup_{w\in G_\mu^*} \frac{((T-T_h)(v), w)_{L^2(D)}}{\|v\|_{L^2}\|w\|_{L^2}} \le c\, h^{\tau_\mu+\tau_\mu^*}.$$

The conclusion follows by applying Theorem 48.1 to Theorem 48.3.

Exercise 48.5 (Nonconforming approximation). The assumption (48.25) holds true for the Laplace operator with homogeneous Dirichlet conditions in a Lipschitz polyhedron with

$$V_{\mathrm{S}} := H^{1+r}(D) \cap V, \qquad r > \frac{1}{2}.$$

After extending a_h to $V_\sharp \times V_\sharp$ by setting $a_\sharp(v,w) := \int_D \nabla_h v \cdot \nabla_h w \,\mathrm{d}x$ where ∇_h is the broken gradient operator (which is an extension of the usual gradient operator to $V_\sharp$; see Definition 36.3 and below), the assumption (48.26) holds true with

$$\|v\|_{V_\sharp}^2 := \sum_{K\in\mathcal{T}_h} \|\nabla v\|_{\boldsymbol{L}^2(K)}^2 + \sum_{K\in\mathcal{T}_h} h_K \|\boldsymbol{n}_K \cdot \nabla v\|_{L^2(\partial K)}^2.$$

Since $T(v) \in H_0^1(D)$ and $S_*(w) \in H_0^1(D)$ for all $v, w \in L^2(D)$, i.e., $\nabla T(v) \in \boldsymbol{L}^2(D)$ and $\nabla S_*(w) \in \boldsymbol{L}^2(D)$, we have

$$\begin{aligned} a_\sharp(T(v), S_*(w)) &= \int_D \nabla_h T(v) \cdot \nabla_h S_*(w) \,\mathrm{d}x \\ &= \int_D \nabla T(v) \cdot \nabla S_*(w) \,\mathrm{d}x = a(T(v), S_*(w)), \end{aligned}$$

for all $v, w \in L^2(D)$. This proves that the assumption (48.27) holds true. Similarly, the assumption (48.28) (related to the restricted and adjoint Galerkin orthogonality properties) holds true because $V_h \cap V \subset V := H_0^1(D)$. The two properties (48.29) are a consequence of the error estimate (36.21). Finally, the best-approximation property (48.30) is a consequence of $V_h \cap V \subset P_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$ and the approximation properties of H^1-conforming finite elements.

Chapter 49

Well-posedness for PDEs in mixed form

Exercises

Exercise 49.1 (Algebraic setting). (i) Derive the counterpart of Theorem 49.12 in the setting of §49.3.1. (*Hint*: assume that the matrix $\mathcal{B}$ has full row rank and consider a basis of $\ker(\mathcal{B})$.) (ii) What happens if the matrix $\mathcal{A}$ is symmetric positive definite?

Exercise 49.2 (Constrained minimization). The goal is to prove Proposition 49.11. (i) Prove that if u minimizes $\mathfrak{E}$ over V_g, there is (a unique) $p \in Q$ such that (u,p) solves (49.35). (*Hint*: proceed as in §49.3.1.) (ii) Prove that (u,p) solves (49.35) if and only if (u,p) is a saddle point of $\mathcal{L}$. (*Hint*: consider $\mathfrak{E}_p : V \to \mathbb{R}$ s.t. $\mathfrak{E}_p(v) := \mathcal{L}(v,p)$ with fixed $p \in Q$.) (iii) Prove that if (u,p) is a saddle point of $\mathcal{L}$, then u minimizes $\mathfrak{E}$ over V_g. (iv) Application: minimize $\mathfrak{E}(v) := 2v_1^2 + 2v_2^2 - 6v_1 + v_2$ over $\mathbb{R}^2$ under the constraint $2v_1 + 3v_2 = -1$.

Exercise 49.3 (Symmetric operator). Let X be a Hilbert space and let $T \in \mathcal{L}(X;X)$ be a bijective symmetric operator. (i) Prove that T^{-1} is symmetric. (ii) Prove that $[\,\lambda \in \sigma(T)\,] \iff [\,\lambda^{-1} \in \sigma(T^{-1})\,]$. (*Hint*: use Corollary 46.18.) (iii) Prove that $\sigma(T) \subset \mathbb{R}$. (*Hint*: consider the sesquilinear form $t_\lambda(x,y) := ((T - \lambda I_X)(x), y)_X$ and use the Lax–Milgram lemma.)

Exercise 49.4 (Sharp stability). The goal is to prove Proposition 49.8. (i) Assume that $\ker(B)$ is nontrivial. Verify that $1 \in \sigma_{\mathrm{p}}(\widetilde{T})$. (ii) Let $\lambda \neq 1$ be in $\sigma(\widetilde{T})$. Prove that $\lambda(\lambda-1) \in \sigma(S)$. (*Hint*: consider the sequence $x_n := (v_n, q_n)$ in X from Corollary 46.18, then observe that $(S(q_n), q_n)_Q = (1-\lambda)^2\langle A(v_n), v_n\rangle_{V',V} + \delta_n$, with $\delta_n := \langle B^*(q_n) + (1-\lambda)A(v_n), A^{-1}B^*(q_n) - (1-\lambda)v_n\rangle_{V',V}$, and prove that $S(q_n) - \lambda(\lambda-1)q_n \to 0$ and $\liminf_{n\to\infty} \|q_n\|_Q > 0$.) (iii) Prove that $\sigma(\widetilde{T}) \subset [\lambda_\sharp^-, \lambda_\flat^-] \cup \{1\} \cup [\lambda_\flat^+, \lambda_\sharp^+]$ with $\lambda_\flat^\pm = \frac{1}{2}(1 \pm (4\frac{\beta^2}{\|a\|} + 1)^{\frac{1}{2}})$, and $\lambda_\sharp^\pm = \frac{1}{2}(1 \pm (4\frac{\|b\|^2}{\alpha} + 1)^{\frac{1}{2}})$. (*Hint*: use Lemma 49.1.) (iv) Conclude. (*Hint*: $\widetilde{T}$ is symmetric with respect to the weighted inner product $(x,y)_{\widetilde{X}} := a(v,w) + (q,r)_Q$.)

Exercise 49.5 (Abstract Helmholtz decomposition). Consider the setting of §49.2 and equip V with the bilinear form a as inner product. (i) Prove that $\mathrm{im}(A^{-1}B^*)$ is closed and that $V = \ker(B) \oplus \mathrm{im}(A^{-1}B^*)$, the sum being a-orthogonal. (*Hint*: use Lemma C.39.) (ii)

Let $f \in \ker(B)^\perp$. Prove that solving $b(v,p) = f(v)$ for all $v \in V$ is equivalent to solving $(S(p),q)_Q = (J_Q^{-1}BA^{-1}(f),q)_Q$ for all $q \in Q$.

Exercise 49.6 (Maxwell's equations). Consider the following problem: For $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, find $\boldsymbol{A}$ and ϕ such that

$$\begin{cases} \nabla\times(\kappa\nabla\times\boldsymbol{A}) + \nu\nabla\phi = \boldsymbol{f}, \\ \nabla\cdot(\nu\boldsymbol{A}) = 0, \\ \boldsymbol{A}_{|\partial D_\mathrm{d}}\times\boldsymbol{n} = \boldsymbol{0},\ \phi_{|\partial D_\mathrm{d}} = 0,\ (\kappa\nabla\times\boldsymbol{A})_{|\partial D_\mathrm{n}}\times\boldsymbol{n} = \boldsymbol{0},\ \boldsymbol{A}_{|\partial D_\mathrm{n}}\cdot\boldsymbol{n} = 0, \end{cases}$$

where κ, ν are real and positive constants (for simplicity), and $|\partial D_\mathrm{d}| > 0$ (see §49.1.3; here we write $\boldsymbol{A}$ in lieu of $\boldsymbol{H}$ and we consider mixed Dirichlet–Neumann conditions). (i) Give a mixed weak formulation of this problem. (*Hint*: use the spaces $\boldsymbol{V}_\mathrm{d} := \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{curl};D) \mid \gamma^\mathrm{c}(\boldsymbol{v})_{|\partial D_\mathrm{d}} = \boldsymbol{0}\}$, where the meaning of the boundary condition is specified in §43.2.1, and $Q_\mathrm{d} := \{q \in H^1(D) \mid \gamma^\mathrm{g}(q)_{|\partial D_\mathrm{d}} = 0\}$.) (ii) Let $B : \boldsymbol{V}_\mathrm{d} \to Q_\mathrm{d}'$ be s.t. $\langle B(\boldsymbol{v}),q\rangle_{Q_\mathrm{d}',Q_\mathrm{d}} := (\nu\boldsymbol{v},\nabla q)_{\boldsymbol{L}^2(D)}$. Let $\boldsymbol{v} \in \ker(B)$. Show that $\nabla\cdot\boldsymbol{v} = 0$ and, if $\boldsymbol{v} \in \boldsymbol{H}^1(D)$, $\gamma^\mathrm{g}(\boldsymbol{v})_{|\partial D_\mathrm{n}}\cdot\boldsymbol{n} = 0$. (*Hint*: recall that ν is constant.) (iii) Accept as a fact that D, ∂D_d, ∂D_n have topological and smoothness properties such that there exists $c > 0$ s.t. $\ell_D\|\nabla\times\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \geq c\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)}$, for all $\boldsymbol{v} \in \ker(B)$, with $\ell_D := \mathrm{diam}(D)$. Show that the above weak problem is well-posed. (*Hint*: use Theorem 49.13.) (iv) Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular sequence of affine meshes. Let $k \geq 0$, let $\boldsymbol{V}_h := \boldsymbol{P}_k^\mathrm{c}(\mathcal{T}_h) \cap \boldsymbol{V}_\mathrm{d}$, and let $Q_h := P_{k+1}^\mathrm{g}(\mathcal{T}_h) \cap Q_\mathrm{d}$. Show that $\nabla Q_h \subset \boldsymbol{V}_h$. (v) Show that the discrete mixed problem is well-posed in $\boldsymbol{V}_h\times Q_h$ assuming that $\partial D_\mathrm{d} = \partial D$. (*Hint*: invoke Theorem 44.6.)

Solution to exercises

Exercise 49.1 (Algebraic setting). (i) Let us consider the linear system (49.27). We have already seen that a necessary condition for the invertibility of the matrix $\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & 0 \end{smallmatrix}\right)$ is that $\mathcal{B}$ has full row rank and that this implies in particular that $M' := N - M \geq 0$. Notice that $M' = \dim(\ker(\mathcal{B}))$ since $\dim(\mathrm{im}(\mathcal{B})) = M$.

If $M' = 0$, i.e., $N = M$, the matrix $\mathcal{B}$ is square and invertible, and it is readily seen that the matrix $\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & 0 \end{smallmatrix}\right)$ is invertible without invoking further assumptions on the matrix $\mathcal{A}$, and we have

$$\begin{pmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & \mathcal{B}^{-1} \\ \mathcal{B}^{-\mathsf{T}} & -\mathcal{B}^{-\mathsf{T}}\mathcal{A}\mathcal{B}^{-1} \end{pmatrix}.$$

Thus, if $N = M$, Theorem 49.12 can be reformulated as follows: the matrix $\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & 0 \end{smallmatrix}\right)$ is invertible iff the matrix $\mathcal{B}$ has full row rank.

Assume now that $N > M$, so that $M' \geq 1$. Let $(\mathsf{J}_i)_{i\in\{1:M'\}}$ be a basis of $\ker(\mathcal{B})$ (recall that by convention the J_i's are column vectors in $\mathbb{R}^N$) and let $\mathcal{J} \in \mathbb{R}^{N\times M'}$ be the rectangular matrix formed by the above basis. Observe that $\ker(\mathcal{J}^\mathsf{T}) = (\ker\mathcal{B})^\perp = \mathrm{im}\,\mathcal{B}^\mathsf{T}$. Recall that $\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & 0 \end{smallmatrix}\right)$ is invertible iff $\ker\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & 0 \end{smallmatrix}\right) = \{0\}$. Assume first that $\ker\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & 0 \end{smallmatrix}\right) = \{0\}$. If $\ker(\mathcal{J}^\mathsf{T}\mathcal{A}\mathcal{J}) \neq \{0\}$, there exists $\mathsf{V} \neq 0$ in $\ker(\mathcal{J}^\mathsf{T}\mathcal{A}\mathcal{J})$. Then, let $\mathsf{U} := \mathcal{J}\mathsf{V}$ and notice that $\mathsf{U} \neq 0$ since the columns of $\mathcal{J}$ are linearly independent. The identity $\ker(\mathcal{J}^\mathsf{T}) = \mathrm{im}\,\mathcal{B}^\mathsf{T}$ implies that there exists P such that $-\mathcal{B}^\mathsf{T}\mathsf{P} = \mathcal{A}\mathcal{J}\mathsf{V} = \mathcal{A}\mathsf{U}$. This is a contradiction since $0 \neq (\mathsf{U},\mathsf{P}) \in \ker\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & 0 \end{smallmatrix}\right)$. Hence, $\mathcal{J}^\mathsf{T}\mathcal{A}\mathcal{J}$ is

invertible. Suppose now that $\mathcal{J}^\mathsf{T}\mathcal{A}\mathcal{J}$ is invertible. Let $(\mathsf{U},\mathsf{P}) \in \mathbb{R}^N \times \mathbb{R}^M$ in $\ker\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & \mathbb{O} \end{smallmatrix}\right)$. Then, $\mathcal{B}\mathsf{U} = 0$, i.e., $\mathsf{U} \in \ker(\mathcal{B})$. Hence, there is $\mathsf{V} \in \mathbb{R}^{M'}$ s.t. $\mathsf{U} = \mathcal{J}\mathsf{V}$. This means that $\mathcal{A}\mathcal{J}\mathsf{V} + \mathcal{B}^\mathsf{T}\mathsf{P} = 0$. Multiplying on the left by $\mathcal{J}^\mathsf{T}$ gives

$$0 = \mathcal{J}^\mathsf{T}\mathcal{A}\mathcal{J}\mathsf{V} + \mathcal{J}^\mathsf{T}\mathcal{B}^\mathsf{T}\mathsf{P} = \mathcal{J}^\mathsf{T}\mathcal{A}\mathcal{J}\mathsf{V} + (\mathcal{B}\mathcal{J})^\mathsf{T}\mathsf{P} = \mathcal{J}^\mathsf{T}\mathcal{A}\mathcal{J}\mathsf{V}.$$

Hence, $\mathcal{J}^\mathsf{T}\mathcal{A}\mathcal{J}\mathsf{V} = 0$, which implies that $\mathsf{V} = 0$, $0 = \mathcal{J}\mathsf{V} = \mathsf{U}$ and $0 = \mathcal{A}\mathcal{J}\mathsf{V} = -\mathcal{B}^\mathsf{T}\mathsf{P}$. But $\ker(\mathcal{B}^\mathsf{T}) = \{0\}$ since $\mathcal{B}$ has full row rank, i.e., $\mathsf{P} = 0$. In conclusion, $\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & \mathbb{O} \end{smallmatrix}\right)$ is invertible.

To sum up, the algebraic counterpart of the operator A_π from Theorem 49.12 is the matrix $\mathcal{A}_\pi = \mathcal{J}^\mathsf{T}\mathcal{A}\mathcal{J} \in \mathbb{R}^{M'\times M'}$, and Theorem 49.12 can be reformulated as follows: the matrix $\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & \mathbb{O} \end{smallmatrix}\right)$ is invertible iff the matrix $\mathcal{B}$ has full row rank and the matrix $\mathcal{A}_\pi$ is invertible.
(ii) If the matrix $\mathcal{A}$ is symmetric positive definite and $N > M$, so is the matrix $\mathcal{A}_\pi$. Therefore, the invertibility of $\left(\begin{smallmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & \mathbb{O} \end{smallmatrix}\right)$ is equivalent to $\mathcal{B}$ having full row rank.

Exercise 49.2 (Constrained minimization). (i) Since $u \in V_g$, $b(u,q) = g(q)$ for all $q \in Q$. Let $B \in \mathcal{L}(V;Q')$ be the operator associated with the bilinear form b. The Euler condition yields $D\mathfrak{E}(u)(h) = 0$ for all $h \in \ker(B)$. Since $D\mathfrak{E}(u)(h) = a(u,h) - f(h)$, we conclude that the linear form $a(u,\cdot) - f(\cdot)$ is in $\ker(B)^\perp = \mathrm{im}(B^*)$ since B is surjective. Hence, there is $p \in Q$ such that $a(u,v) + b(v,p) = f(v)$ for all $v \in V$. The uniqueness of p follows from the injectivity of B^*.
(ii) Assume that (u,p) solves (49.35). Then $\mathcal{L}(u,q) = \mathcal{L}(u,p)$ for all $q \in Q$. Moreover, the functional $\mathfrak{E}_p : V \to \mathbb{R}$ s.t. $\mathfrak{E}_p(v) := \mathcal{L}(v,p)$ (with p fixed) is strictly convex and $D\mathfrak{E}_p(u)(h) = 0$ for all $h \in V$ since $D\mathfrak{E}_p(u)(h) = a(u,h) + b(h,p) - f(h)$. This implies that u minimizes $\mathfrak{E}_p$ over V. Conversely, assume that (u,p) is a saddle point of $\mathcal{L}$. This implies that $\mathcal{L}(u,q) - \mathcal{L}(u,p) = \langle B(u) - g, q - p\rangle_{Q',Q} \le 0$ for all $q \in P$. Hence, $\langle B(u) - g, q\rangle_{Q',Q} \le 0$, and taking $\pm q$, we infer that $\langle B(u) - g, q\rangle_{Q',Q} = 0$ for all $q \in Q$. Therefore, $B(u) = g$. Moreover, u minimizes the functional $\mathfrak{E}_p$ over p, whence we infer that $D\mathfrak{E}_p(u)(h) = 0$ for all $h \in V$, i.e., $a(u,h) + b(h,p) - f(h) = 0$ for all $h \in V$.
(iii) Assume that (u,p) is a saddle point of $\mathcal{L}$. We have already seen that the left inequality in (49.29) implies that $B(u) = g$, i.e., $u \in V_g$. Moreover, taking $v \in V_g$, we can see from the right inequality in (49.29) that $\mathfrak{E}(u) - \mathfrak{E}(v) = \mathcal{L}(u,p) - \mathcal{L}(v,p) \le 0$.
(iv) Using the above results, we infer that u is the minimizer of $\mathfrak{E}$ over $\mathbb{R}^2$ under the above constraint if and only if $(u,p) := (u_1,u_2,p)$ is a saddle point of $\mathcal{L}(v,q) = \mathfrak{E}(v) + q(2v_1 + 3v_2 + 1)$. The optimality conditions are

$$\begin{aligned} 0 &= \partial_{v_1}\mathcal{L}(u,p) = 4u_1 - 6 + 2p, \\ 0 &= \partial_{v_2}\mathcal{L}(u,p) = 4u_2 + 1 + 3p, \\ 0 &= \partial_q\mathcal{L}(u,p) = 2u_1 + 3u_2 + 1. \end{aligned}$$

The solution to this linear system is $(u_1,u_2,p) = (1,-1,1)$. Hence, the minimizer is $u = (1,-1)$, and the minimum is $\mathfrak{E}(u) = -3$.

Exercise 49.3 (Symmetric operator). (i) Using the symmetry of T, we infer that for all $x, y \in X$,

$$(T^{-1}(x),y)_X = (T^{-1}(x),TT^{-1}(y))_X = (T^{-1}(y),TT^{-1}(x))_X = (T^{-1}(y),x)_X.$$

(ii) Let $\lambda \in \sigma(T)$. Owing to Corollary 46.18, there is $(x_n)_{n\in\mathbb{N}}$ in X such that $\|v_n\|_X = 1$ for all $n \in \mathbb{N}$ and $T(v_n) - \lambda v_n \to 0$ as $n \to \infty$. Since T is bijective, $\lambda \neq 0$ so that $\lambda^{-1}T^{-1}$ is bounded. This implies that $\lambda^{-1}v_n - T^{-1}(v_n) \to 0$ as $n \to \infty$, which shows that $\lambda^{-1} \in \sigma(T^{-1})$. The proof

for the converse is identical.
(iii) Assume that $\lambda = \alpha + \mathrm{i}\beta \in \sigma(T)$ with $\beta \neq 0$. The sesquilinear form $t_\lambda(x,y) := ((T - \lambda I_X)(x), y)_X$ is bounded and coercive, this latter property following from

$$\Re\big(-\mathrm{i}t_\lambda(x,x)\big) = \beta\|x\|_X^2, \qquad \forall x \in X.$$

Hence, for all $y \in X$, there is a unique $x \in X$ such that $t_\lambda(x,z) = (y,z)_X$ for all $z \in X$ owing to the Lax–Milgram lemma. This implies that $(T - \lambda I_X)(x) = y$ showing that $T - \lambda I_X$ is bijective. Hence, $\lambda \notin \sigma(T)$.

Exercise 49.4 (Sharp stability). (i) Let $v \in \ker(B) \setminus \{0\}$. Then $\widetilde{T}(v,0) = (v,0)$, so that $1 \in \sigma_{\mathrm{p}}(\widetilde{T})$.
(ii) Consider a sequence $(x_n)_{n\in\mathbb{N}}$ in X such that $\|x_n\|_X = 1$ for all $n \in \mathbb{N}$ and $\widetilde{T}(x_n) - \lambda x_n \to 0$ as $n \to \infty$. Writing $x_n := (v_n, q_n)$, we infer that $(1-\lambda)v_n + A^{-1}B^*(q_n) \to 0$ and $J_Q^{-1}B(v_n) - \lambda q_n \to 0$. This implies that $S(q_n) - \lambda(\lambda-1)q_n \to 0$. We observe that

$$(S(q_n), q_n)_Q = (1-\lambda)^2\langle A(v_n), v_n\rangle_{V',V} + \delta_n,$$

with

$$\delta_n := \langle B^*(q_n) + (1-\lambda)A(v_n), A^{-1}B^*(q_n) - (1-\lambda)v_n\rangle_{V',V},$$

and $\delta_n \to 0$ since $B^*(q_n) + (1-\lambda)A(v_n) \to 0$ (since A is bounded) and $A^{-1}B^*(q_n) - (1-\lambda)v_n$ is bounded in V (since x_n is bounded in X). Owing to the coercivity of A and the characterization of $\sigma(S)$, we infer that

$$\begin{aligned}\left(\frac{\|b\|^2}{\alpha(1-\lambda)^2} + 1\right)\|q_n\|_Q^2 &\geq \frac{1}{\alpha(1-\lambda)^2}(S(q_n), q_n)_Q + \|q_n\|_Q^2 \\ &\geq \|x_n\|_X^2 + \frac{1}{\alpha(1-\lambda)^2}\delta_n.\end{aligned}$$

This shows that $\liminf_{n\to\infty}\|q_n\|_Q > 0$. Recalling that $S(q_n) - \lambda(\lambda-1)q_n \to 0$, we conclude that $\lambda(\lambda-1) \in \sigma(S)$.
(iii) Lemma 49.1 implies that $\lambda(\lambda-1) \in \big[\frac{\beta^2}{\|a\|}, \frac{\|b\|^2}{\alpha}\big]$. A simple reasoning on the quadratic function $\lambda \mapsto \lambda(\lambda-1)$ leads to the expected result on $\sigma(\widetilde{T})$, recalling that $1 \in \sigma(\widetilde{T})$.
(iv) We observe that $\widetilde{T}$ is symmetric with respect to the weighted inner product $(x,y)_{\widetilde{X}} := a(v,w) + (q,r)_Q$. Let $\|\cdot\|_{\widetilde{X}}$ be the induced norm in X. Equipping X with this norm, we infer that

$$\begin{aligned}\|\widetilde{T}\|_{\mathcal{L}(X;X)} &= \sup_{\lambda\in\sigma(\widetilde{T})}|\lambda| = \lambda_\sharp^+, \\ \|\widetilde{T}^{-1}\|_{\mathcal{L}(X;X)} &= \sup_{\lambda\in\sigma(\widetilde{T})}|\lambda|^{-1} = (-\lambda_\flat^-)^{-1}.\end{aligned}$$

Since $\widetilde{T}(x) = y$ with $y := (A^{-1}(f), J_Q^{-1}(g))$ whenever $x := (u,p)$ solves (49.35), we conclude that (49.26) holds true.

Exercise 49.5 (Abstract Helmholtz decomposition). (i) Since

$$\|a\|\|A^{-1}B^*(q)\|_V^2 \geq \langle B^*(q), A^{-1}B^*(q)\rangle_{V',V} = (S(q),q)_Q \geq \frac{\beta^2}{\|a\|}\|q\|_Q^2$$

owing to Lemma 49.1, Lemma C.39 implies that $\mathrm{im}(A^{-1}B^*)$ is closed. That $V = \ker(B) + \mathrm{im}(A^{-1}B^*)$ results from the fact that the saddle point problem (49.33) has a solution with right-hand side $(f,g) = (A(v),0)$ for all $v \in V$. Finally, let $v_0 \in \ker(B)$ and $v_1 \in \mathrm{im}(A^{-1}B^*)$ so that $v_1 = A^{-1}B^*(q)$ for some $q \in Q$. Hence,

$$a(v_0,v_1) = \langle A(v_0), A^{-1}B^*(q)\rangle_{V',V} = \langle B(v_0), q\rangle_{Q',Q} = 0.$$

This proves the a-orthogonality between $\ker(B)$ and $\mathrm{im}(A^{-1}B^*)$.
(ii) Solving $b(v,p) = f(v)$ for all $v \in V$ amounts to $B^*(p) = f$ in V'. Since both forms vanish on $\ker(B)$, it is enough to assert that $\langle B^*(p), v_1\rangle_{V',V} = \langle f, v_1\rangle_{V',V}$ for all $v_1 \in \mathrm{im}(A^{-1}B^*)$. Therefore, we have for all $q \in Q$,

$$(S(p),q)_Q = \langle BA^{-1}B^*(p), q\rangle_{Q',Q} = \langle B^*(p), A^{-1}B^*(q)\rangle_{V',V} = (f, A^{-1}B^*(q))_{V',V}.$$

This proves the equivalence.

Exercise 49.6 (Maxwell's equations). (i) We obtain a weak formulation of the problem by testing the equations with smooth vector fields $\boldsymbol{v}$ and smooth scalar fields q (recall that we can work here with real-valued functions and fields since ν and κ are real numbers):

$$\begin{cases} \displaystyle\int_D (\kappa\nabla\times\boldsymbol{A}\cdot\nabla\times\boldsymbol{v})\,\mathrm{d}x - \int_{\partial D}(\kappa(\nabla\times\boldsymbol{A})\times\boldsymbol{n})\cdot\boldsymbol{v}\,\mathrm{d}s + \int_D \nu\boldsymbol{v}\cdot\nabla\phi\,\mathrm{d}x = \int_D \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x, \\ \displaystyle\int_D \nu\boldsymbol{A}\cdot\nabla q\,\mathrm{d}x - \int_{\partial D}(\nu\boldsymbol{A}\cdot\boldsymbol{n})q\,\mathrm{d}s = 0. \end{cases}$$

We now apply the boundary conditions assuming that the test functions satisfy $\boldsymbol{v}_{|\partial D_\mathrm{d}}\times\boldsymbol{n} = 0$ and $q_{|\partial D_\mathrm{d}} = 0$, which leads to

$$\begin{cases} \displaystyle\int_D (\kappa\nabla\times\boldsymbol{A}\cdot\nabla\times\boldsymbol{v})\,\mathrm{d}x + \int_D \nu\boldsymbol{v}\cdot\nabla\phi\,\mathrm{d}x = \int_D \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x, \\ \displaystyle\int_D \nu\boldsymbol{A}\cdot\nabla q\,\mathrm{d}x = 0. \end{cases}$$

We can make sense of the above informal argument by assuming $\boldsymbol{A},\boldsymbol{v} \in \boldsymbol{V}_\mathrm{d}$ and $\phi, q \in Q_\mathrm{d}$ where

$$\begin{aligned} \boldsymbol{V}_\mathrm{d} &:= \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{curl};D) \mid \gamma^\mathrm{c}(\boldsymbol{v})_{|\partial D_\mathrm{d}} = \boldsymbol{0}\}, \\ Q_\mathrm{d} &:= \{q \in H^1(D) \mid \gamma^\mathrm{g}(q)_{|\partial D_\mathrm{d}} = 0\}, \end{aligned}$$

where the boundary condition in $\boldsymbol{V}_\mathrm{d}$ means that $\int_D(\boldsymbol{v}\cdot\nabla\times\boldsymbol{w} - (\nabla\times\boldsymbol{v})\cdot\boldsymbol{w})\,\mathrm{d}x = 0$ for all $\boldsymbol{w} \in \boldsymbol{H}^1(D)$ s.t. $\gamma^\mathrm{g}(\boldsymbol{w})_{|\partial D_\mathrm{d}} \in \widetilde{\boldsymbol{H}}^{\frac12}(\partial D_\mathrm{d})$. We equip $\boldsymbol{V}_\mathrm{d}$ with the norm of $\boldsymbol{H}(\mathrm{curl};D)$ (see the proof of Theorem 43.1) and Q_d with the norm of $H^1(D)$. Then, $\boldsymbol{V}_\mathrm{d}$ is a closed subspace of $\boldsymbol{H}(\mathrm{curl};D)$ and Q_d is a closed subspace of $H^1(D)$. We introduce the bilinear forms $a(\boldsymbol{A},\boldsymbol{v}) := \int_D(\kappa\nabla\times\boldsymbol{A}\cdot\nabla\times\boldsymbol{v})\,\mathrm{d}x$ and $b(\boldsymbol{A},q) := \int_D \nu\boldsymbol{A}\cdot\nabla q\,\mathrm{d}x$, and the linear form $\ell(\boldsymbol{v}) := \int_D \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x$. The above problem is reformulated as follows: Find $\boldsymbol{A} \in \boldsymbol{V}_\mathrm{d}$ and $\phi \in Q_\mathrm{d}$ such that

$$\begin{aligned} a(\boldsymbol{A},\boldsymbol{v}) + b(\boldsymbol{v},\phi) &= \ell(\boldsymbol{v}) \quad && \forall \boldsymbol{v} \in \boldsymbol{V}_\mathrm{d}, \\ b(\boldsymbol{A},q) &= 0 && \forall q \in Q_\mathrm{d}. \end{aligned}$$

(ii) Let $\boldsymbol{v} \in \ker(B)$, i.e., $0 = \int_D \nu\boldsymbol{v}\cdot\nabla q\,\mathrm{d}x$ for all $q \in Q_\mathrm{d}$. Taking q arbitrary in $C_0^\infty(D)$ and since ν is constant, we infer that $\nabla\cdot\boldsymbol{v} = 0$ in D. If $\boldsymbol{v} \in \boldsymbol{H}^1(D)$, we infer that for all $\phi \in \widetilde{H}^{\frac12}(\partial D_\mathrm{n})$, we have

$$\int_{\partial D_\mathrm{n}}(\boldsymbol{v}\cdot\boldsymbol{n})\phi\,\mathrm{d}s = \int_{\partial D}(\boldsymbol{v}\cdot\boldsymbol{n})\widetilde{\phi}\,\mathrm{d}s = \int_D \big(\boldsymbol{v}\cdot\nabla l(\widetilde{\phi}) + (\nabla\cdot\boldsymbol{v})l(\widetilde{\phi})\big)\,\mathrm{d}x = 0 + 0 = 0,$$

where $\widetilde{\phi}$ is the zero extension of ϕ to ∂D and $l(\widetilde{\phi})$ is a lifting of $\widetilde{\phi}$ in $Q_{\rm d}$. The above identity implies that $\gamma^{\rm g}(\boldsymbol{v})_{|\partial D_{\rm n}}\cdot\boldsymbol{n}=0$.
(iii) We are going to use the Babuška–Brezzi theorem to prove the well-posedness of the mixed formulation, that is, we have to prove (49.36) and (49.37). Since we have $\|\nabla\times\boldsymbol{v}\|^2_{\boldsymbol{L}^2(D)}\geq \ell_D^{-2}c^2\|\boldsymbol{v}\|^2_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{v}\in\ker(B)$, we infer that

$$\|\nabla\times\boldsymbol{v}\|^2_{\boldsymbol{L}^2(D)}\geq\frac{c^2}{1+c^2}(\|\nabla\times\boldsymbol{v}\|^2_{\boldsymbol{L}^2(D)}+\ell_D^{-2}\|\boldsymbol{v}\|^2_{\boldsymbol{L}^2(D)})=\frac{c^2\ell_D^{-2}}{1+c^2}\|\boldsymbol{v}\|^2_{\boldsymbol{H}(\mathrm{curl};D)}.$$

This shows that $a(\boldsymbol{v},\boldsymbol{v})\geq\frac{c^2\ell_D^{-2}\kappa}{1+c^2}\|\boldsymbol{v}\|^2_{\boldsymbol{H}(\mathrm{curl};D)}$ for all $\boldsymbol{v}\in\ker(B)$, which proves (49.36).
Since $|\partial D_{\rm d}|>0$, we equip $Q_{\rm d}$ with the norm $\|q\|_{Q_{\rm d}}:=|q|_{H^1(D)}$. Let q be a nonzero member of $Q_{\rm d}$. Letting $\boldsymbol{v}_q:=\nu^{-1}\nabla q$, we verify that $\gamma^{\rm c}(\boldsymbol{v}_q)_{|\partial D_{\rm d}}=\boldsymbol{0}$ and $\nu\|\boldsymbol{v}_q\|_{\boldsymbol{H}(\mathrm{curl};D)}\leq|q|_{H^1(D)}=\|q\|_{Q_{\rm d}}$ (note that $\boldsymbol{v}_q$ is curl-free since ν is constant). The definition of $\boldsymbol{v}_q$ implies that $b(\boldsymbol{v}_q,q)=\int_D\nu\boldsymbol{v}_q\cdot\nabla q\,\mathrm{d}x=\|\nabla q\|^2_{\boldsymbol{L}^2(D)}=\|q\|^2_{Q_{\rm d}}$, which, in turn, gives

$$\sup_{\boldsymbol{w}\in\boldsymbol{V}_{\rm d}}\frac{|b(\boldsymbol{w},q)|}{\|\boldsymbol{w}\|_{\boldsymbol{H}(\mathrm{curl};D)}}\geq\frac{|b(\boldsymbol{v}_q,q)|}{\|\boldsymbol{v}_q\|_{\boldsymbol{H}(\mathrm{curl};D)}}=\frac{\|q\|^2_{Q_{\rm d}}}{\|\boldsymbol{v}_q\|_{\boldsymbol{H}(\mathrm{curl};D)}}\geq\nu\|q\|_{Q_{\rm d}}.$$

This proves (49.37). In conclusion, the weak mixed problem is well-posed.
(iv) Recall that

$$\begin{aligned}\boldsymbol{V}_h&=\{\boldsymbol{v}_h\in\boldsymbol{H}(\mathrm{curl};D)\mid\boldsymbol{\psi}_K^{\rm c}(\boldsymbol{v}_{h|K})\in\mathfrak{N}_{k,3},\forall K\in\mathcal{T}_h;\ \boldsymbol{v}_{h|\partial D_{\rm d}}\times\boldsymbol{n}=\boldsymbol{0}\},\\ Q_h&=\{q_h\in H^1(D)\mid\psi_K^{\rm g}(q_{h|K})\in\widehat{P},\forall K\in\mathcal{T}_h;\ q_{h|\partial D_{\rm d}}=0\}.\end{aligned}$$

Here, $\widehat{P}$ is either $\mathbb{P}_{k+1,3}$ or $\mathbb{Q}_{k+1,3}$ depending on the shape of the cells. Recalling the commuting properties stated in Lemma 16.16, we have $\nabla q_h\in\boldsymbol{H}(\mathrm{curl};D)$ and $\boldsymbol{\psi}_K^{\rm c}(\nabla q_{h|K})\in\mathfrak{N}_{k,3}$ for all $q_h\in Q_h$. Moreover, the boundary condition $q_{h|\partial D_{\rm d}}=0$ implies that $\nabla q_{h|\partial D_{\rm d}}\times\boldsymbol{n}=\boldsymbol{0}$ for all $q_h\in Q_h$. Hence, $\nabla Q_h\subset\boldsymbol{V}_h$.
(v) The discrete mixed problem posed in $\boldsymbol{V}_h\times Q_h$ consists of seeking $\boldsymbol{A}_h\in\boldsymbol{V}_h$ and $\phi_h\in Q_h$ such that

$$\begin{aligned}&a(\boldsymbol{A}_h,\boldsymbol{v}_h)+b(\boldsymbol{v}_h,\phi_h)=\ell(\boldsymbol{v}_h),&&\forall\boldsymbol{v}_h\in\boldsymbol{V}_h,\\ &b(\boldsymbol{A}_h,q_h)=0&&\forall q_h\in Q_h.\end{aligned}$$

Proving the well-posedness of this discrete problem can be done by proving that a is coercive on the discrete space

$$\ker(B_h):=\{\boldsymbol{v}_h\in\boldsymbol{V}_h\mid b(\boldsymbol{v}_h,q_h)=0,\forall q_h\in Q_h\}.$$

Since we have assumed that $\partial D_{\rm d}=\partial D$, this is exactly the coercivity statement made in Theorem 44.6 under the form of a discrete Poincaré–Steklov inequality, i.e., a is coercive on $\ker(B_h)$ with a coercivity constant that is uniform w.r.t. the mesh size. We also need to prove that the discrete counterpart of the inf-sup condition (49.37) holds true. Let q_h be a nonzero member of Q_h. Letting $\boldsymbol{v}_h:=\nu^{-1}\nabla q_h$, we have already verified that $\boldsymbol{v}_h\in\boldsymbol{V}_h$. Moreover, $\nu\|\boldsymbol{v}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}\leq\|q_h\|_{Q_{\rm d}}$, and proceeding as in Step (iv), we infer that

$$\sup_{\boldsymbol{w}_h\in\boldsymbol{V}_h}\frac{|b(\boldsymbol{w}_h,q_h)|}{\|\boldsymbol{w}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}}\geq\nu\|q_h\|_{H^1(D)}.$$

In conclusion, the discrete mixed problem is well-posed.

Chapter 50

Mixed finite element approximation

Exercises

Exercise 50.1 (Algebraic setting). Let $\mathcal{A} := \left(\begin{smallmatrix} 1 & \sqrt{2} \\ \frac{1}{\sqrt{2}} & 0 \end{smallmatrix}\right)$ and $\mathcal{B} := (1,0)^\mathsf{T}$. Show that

$$\inf_{\mathsf{V}\in\ker(\mathcal{B})} \sup_{\mathsf{W}\in\ker(\mathcal{B})} \frac{\mathsf{W}^\mathsf{T}\mathcal{A}\mathsf{V}}{\|\mathsf{W}\|_{\ell^2(\mathbb{R}^2)}\|\mathsf{V}\|_{\ell^2(\mathbb{R}^2)}} < \inf_{\mathsf{V}\in\mathbb{R}^2} \sup_{\mathsf{W}\in\mathbb{R}^2} \frac{\mathsf{W}^\mathsf{T}\mathcal{A}\mathsf{V}}{\|\mathsf{W}\|_{\ell^2(\mathbb{R}^2)}\|\mathsf{V}\|_{\ell^2(\mathbb{R}^2)}}.$$

(*Hint*: one number is equal to 0 and the other is equal to 1.)

Exercise 50.2 (Saddle point problem). Let V, Q be Hilbert spaces and let a be a symmetric, coercive, bilinear form. Consider the discrete problem (50.2) and the bilinear form $t(y,z) := a(v,w) + b(w,q) + b(v,r)$ for all $y := (v,q), z := (w,r) \in X := V\times Q$. Let $X_h := V_h\times Q_h$ and consider the linear map $P_h \in \mathcal{L}(X;X_h)$ such that for all $x \in X$, $P_h(x) \in X_h$ is the unique solution of $t(P_h(x), y_h) = t(x, y_h)$ for all $y_h \in X_h$. Equip X and X_h with the norm $\|(v,q)\|_{\widetilde{X}} := (\|v\|_a^2 + \|q\|_Q^2)^{\frac12}$ with $\|v\|_a^2 := a(v,v)$. (i) Prove that $\|P_h\|_{\mathcal{L}(X;X)} \le \tilde{c}_h := \frac{(4\frac{\|b\|^2}{\alpha}+1)^{\frac12}+1}{(4\frac{\beta_h^2}{\|a\|}+1)^{\frac12}-1}$. (*Hint*: use Proposition 49.8.) (ii) Prove that $\|u-u_h\|_a^2 + \|p-p_h\|_Q^2 \le \tilde{c}_h^2(\inf_{v_h\in V_h}\|u-u_h\|_a^2 + \inf_{q_h\in Q_h}\|p-q_h\|_Q^2)$. (*Hint*: see the proof of Theorem 5.14.)

Exercise 50.3 (Error estimate). (i) Prove directly the estimate (50.7a) with c'_{1h} replaced by $c''_{1h} := (1+\frac{\|a\|}{\alpha_h})(1+\frac{\|b\|}{\beta_h})$. (*Hint*: consider $z_h \in V_h$ s.t. $B_h(z_h) := B_h(u_h - v_h)$ with $v_h \in V_h$ arbitrary.) (ii) Assume that V is a Hilbert space, $\ker(B_h) \subset \ker(B)$, and $g := 0$. Prove that $\|u-u_h\|_V \le \frac{\|a\|}{\alpha_h}\inf_{v_h\in\ker(B_h)}\|u-v_h\|_V$.

Exercise 50.4 (Bound on $\mathcal{A}$ and $\mathcal{B}$). (i) Prove Proposition 50.12. (*Hint*: observe that $(\mathcal{A}\mathsf{U})^\mathsf{T}\mathsf{Y} = a(\mathsf{R}_\varphi(\mathsf{U}),\mathsf{R}_\varphi(\mathsf{Y}))$.) (ii) Let $\mathcal{J}_V \in \mathbb{R}^{N\times N}$ be the symmetric positive definite matrix with entries $\mathcal{J}_{V,ij} := (\varphi_i,\varphi_j)_X$ for all $i,j\in\{1{:}N\}$. Let $\|\cdot\|_{\ell^2(\mathbb{R}^N)}$ denote the Euclidean norm in $\mathbb{R}^N$. Verify that $\|\mathsf{R}_\varphi(\mathsf{U})\|_V = \|\mathcal{J}_V^{\frac12}\mathsf{U}\|_{\ell^2(\mathbb{R}^N)}$ and $\|\mathsf{U}\|_{\ell^2(\mathbb{R}^N)} = \|\mathcal{J}_V^{-\frac12}\mathsf{U}\|_{\ell^2(\mathbb{R}^N)}$ for all $\mathsf{U}\in\mathbb{R}^N$.

Exercise 50.5 ($\mathcal{S}_\rho$). The goal is to prove the identity (50.17). (i) Verify that $\mathcal{A}_\rho^{-1} = \mathcal{A}^{-1} - \rho\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}(\mathcal{M}_Q + \rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1}$. (*Hint*: multiply the right-hand side by $\mathcal{A}_\rho$ and develop the product.) (ii) Infer that $\mathcal{S}_\rho = \mathcal{S} - \rho\mathcal{S}(\mathcal{M}_Q + \rho\mathcal{S})^{-1}\mathcal{S}$. (iii) Conclude. (*Hint*: multiply the right-hand side by $\rho\mathcal{M}_Q^{-1} + \mathcal{S}^{-1}$.)

Exercise 50.6 (Penalty). (i) Prove Proposition 50.18. (*Hint*: verify that $\mathcal{C}(\mathsf{U} - \mathsf{U}_\epsilon, \mathsf{P} - \mathsf{P}_\epsilon)^\mathsf{T} = (0, -\epsilon\mathcal{M}_Q\mathsf{P}_\epsilon)^\mathsf{T}$ and use Proposition 50.12.) (ii) Replace the mass matrix $\mathcal{M}_Q$ by the identity matrix $\mathcal{I}_M$ times a positive coefficient λ in (50.18). Does the method still converge? Is there any interest of doing so? Can you think of another choice?

Exercise 50.7 (Inexact Minres and DPG). Let V, Y be Hilbert spaces and $B \in \mathcal{L}(V; Y')$ be s.t. $\beta\|v\|_V \le \|B(v)\|_{Y'} \le \|b\|\|v\|_V$ for all $v \in V$ with $0 < \beta \le \|b\| < \infty$. Set $b(v, y) := \langle B(v), y\rangle_{Y',Y}$. Let $f \in Y'$. Let $J_Y : Y \to Y'$ denote the isometric Riesz–Fréchet isomorphism. (i) Show that the Minres problem $\min_{v\in V}\|f - B(v)\|_{Y'}$ has a unique solution $u \in V$. (*Hint*: introduce the sesquilinear form $a(v, w) := \langle B(v), J_Y^{-1}(B(w))\rangle_{Y',Y}$ and invoke the Lax–Milgram Lemma.) (ii) Let $\{V_h \subset V\}_{h\in\mathcal{H}}$ and $\{Y_h \subset Y\}_{h\in\mathcal{H}}$ be sequences of subspaces approximating V and Y, respectively. Assume that there is $\beta_0 > 0$ s.t. for all $h \in \mathcal{H}$,

$$\inf_{v_h\in V_h}\sup_{y_h\in Y_h}\frac{|b(v_h, y_h)|}{\|v_h\|_V\|y_h\|_Y} \ge \beta_0. \tag{50.1}$$

Let $I_h : Y_h \to Y$ be the canonical injection and $I_h^* : Y' \to Y_h'$. Show that the inexact Minres problem $\min_{v_h\in V_h}\|I_h^*(f - B(v_h))\|_{Y_h'}$ has a unique solution $u_h \in V_h$. (*Hint*: introduce the residual representative $r_h := J_{Y_h}^{-1}I_h^*(f - B(u_h)) \in V_h$ and show that the pair $(u_h, r_h) \in V_h \times Y_h$ solves a saddle point problem.) (iii) Show that the residual representative $r_h \in Y_h$ is the unique solution of the following constrained minimization problem: $\min_{z_h\in Y_h\cap(I_h^*(B(V_h)))^\perp}\frac{1}{2}\|z_h\|_Y^2 - \langle I_h^*(f), z_h\rangle_{Y_h',Y_h}$. (*Hint*: see Proposition 49.11.) (iv) Assume now that $f \in \mathrm{im}(B)$ so that $B(u) = f$. Prove that there is c s.t. $\|u - u_h\|_V \le c\inf_{w_h\in V_h}\|u - w_h\|_V$ for all $h \in \mathcal{H}$. (*Hint*: use a Fortin operator.) *Note*: since $\beta\|v_h\|_V \le \|B(v_h)\|_{Y'}$ for all $v_h \in V_h$, it is natural to expect that the inf-sup condition (50.1) is satisfied if the subspace $Y_h \subset Y$ is chosen rich enough. The inexact residual minimization in a discrete dual norm is at the heart of the discontinuous Petrov–Galerkin (dPG) method; see Demkowicz and Gopalakrishnan [14], Gopalakrishnan and Qiu [18], Carstensen et al. [10]. The extension to reflexive Banach spaces is studied in Muga and van der Zee [35].

Solution to exercises

Exercise 50.1 (Algebraic setting). Since $\mathcal{A}$ is symmetric, we have

$$\inf_{\mathsf{V}\in\mathbb{R}^2}\sup_{\mathsf{W}\in\mathbb{R}^2}\frac{\mathsf{W}^\mathsf{T}\mathcal{A}\mathsf{V}}{\|\mathsf{W}\|_{\ell^2(\mathbb{R}^2)}\|\mathsf{V}\|_{\ell^2(\mathbb{R}^2)}} = |\lambda_{\min}(\mathcal{A})|,$$

where $\lambda_{\min}(\mathcal{A})$ is the eigenvalue of $\mathcal{A}$ with the smallest absolute value. A simple computation shows that the eigenvalues of $\mathcal{A}$ are -1 and 2, so that $|\lambda_{\min}(\mathcal{A})| = 1$. Moreover, $\ker\mathcal{B} = \mathrm{span}\{e_2\}$ with $e_2 := (0, 1)^\mathsf{T}$. But

$$e_2^\mathsf{T}\mathcal{A}e_2 = (0, 1)^\mathsf{T}(\sqrt{2}, 0) = 0.$$

Hence, we have

$$\inf_{\mathsf{V}\in\ker(\mathcal{B})}\sup_{\mathsf{W}\in\ker(\mathcal{B})}\frac{\mathsf{W}^\mathsf{T}\mathcal{A}\mathsf{V}}{\|\mathsf{W}\|_{\ell^2(\mathbb{R}^2)}\|\mathsf{V}\|_{\ell^2(\mathbb{R}^2)}} = 0.$$

Exercise 50.2 (Saddle point problem). (i) Owing to Proposition 49.8, we infer that

$$\inf_{y_h\in X_h}\sup_{z_h\in X_h}\frac{t(y_h,z_h)}{\|y_h\|_{\widetilde{X}}\|z_h\|_{\widetilde{X}}}=\frac{(4\frac{\beta_h^2}{\|a\|}+1)^{\frac12}-1}{2},$$
$$\sup_{y\in X}\sup_{z\in X}\frac{t(y,z)}{\|y\|_{\widetilde{X}}\|z\|_{\widetilde{X}}}=\frac{(4\frac{\|b\|^2}{\alpha}+1)^{\frac12}+1}{2}.$$

This implies the bound on $\|P_h\|_{\mathcal{L}(X;X)}$.
(ii) Observing that X_h is pointwise invariant under P_h, we infer that

$$\|x-x_h\|_{\widetilde{X}}=\|(I-P_h)(x)\|_{\widetilde{X}}=\|(I-P_h)(x-y_h)\|_{\widetilde{X}}\le\|I-P_h\|_{\mathcal{L}(X;X)}\|x-y_h\|_{\widetilde{X}},$$

for all $y_h\in X_h$. We conclude observing that $\|I-P_h\|_{\mathcal{L}(X;X)}=\|P_h\|_{\mathcal{L}(X;X)}$.

Exercise 50.3 (Velocity estimate). (i) Let $v_h\in V_h$. Owing to the surjectivity of the operator B_h implied by the inf-sup condition (50.4b), there exists $z_h\in V_h$ such that $B_h(z_h)=B_h(u_h-v_h)$ and $\beta_h\|z_h\|_V\le\|B_h(u_h-v_h)\|_{Q_h'}$. Since

$$\begin{aligned}\|B_h(u_h-v_h)\|_{Q_h'}&\le\sup_{q_h\in Q_h}\frac{|b(u_h-v_h,q_h)|}{\|q_h\|_Q}\\&=\sup_{q_h\in Q_h}\frac{|b(u-v_h,q_h)|}{\|q_h\|_Q}\\&\le\|b\|\,\|u-v_h\|_V,\end{aligned}$$

where we used the Galerkin orthogonality property for the second equation in (50.2) (i.e., $b(u-u_h,q_h)=0$ for all $q_h\in Q_h$), we infer that

$$\beta_h\|z_h\|_V\le\|b\|\,\|u-v_h\|_V.$$

Let us set $w_h:=v_h+z_h$. Since $u_h-w_h\in\ker(B_h)$, we infer from the inf-sup condition (50.4a) that

$$\begin{aligned}\alpha_h\|u_h-w_h\|_V&\le\sup_{y_h\in\ker(B_h)}\frac{|a(u_h-w_h,y_h)|}{\|y_h\|_V}\\&=\sup_{y_h\in\ker(B_h)}\frac{|a(u_h-u,y_h)+a(u-w_h,y_h)|}{\|y_h\|_V}\\&=\sup_{y_h\in\ker(B_h)}\frac{|b(y_h,p-p_h)+a(u-w_h,y_h)|}{\|y_h\|_V},\end{aligned}$$

where we used the Galerkin orthogonality property for the first equation in (50.2). The rest of the proof is identical to that of Lemma 50.2.
(ii) Set $V_0:=\ker(B)$. Let $P_h:V_0\to\ker(B_h)$ mapping $u\in V_0$ to the unique solution $u_h\in\ker(B_h)$ of $a(u_h-u,w_h)=0$ for all $w_h\in\ker(B_h)$. Then $\ker(B_h)$ is pointwise invariant under P_h, and $\|P_h\|_{\mathcal{L}(V_0;V_0)}\le\frac{\|a\|}{\alpha_h}$. To conclude, we observe that

$$\|u-u_h\|_V=\|(I-P_h)(u)\|_V=\|(I-P_h)(u-v_h)\|_V\le\|I-P_h\|_{\mathcal{L}(V_0;V_0)}\|u-v_h\|_V,$$

for all $v_h\in\ker(B_h)$ and that $\|I-P_h\|_{\mathcal{L}(V_0;V_0)}=\|P_h\|_{\mathcal{L}(V_0;V_0)}$ (see the proof of Theorem 5.14).

Exercise 50.4 (Bound on $\mathcal{A}$ and $\mathcal{B}$). (i) Let $\mathsf{U},\mathsf{Y}\in\mathbb{R}^N$ and set $u_h:=\mathsf{R}_\varphi(\mathsf{U})$ and $y_h:=\mathsf{R}_\varphi(\mathsf{Y})$. Using the hint, we infer that $\frac{(\mathcal{A}\mathsf{U})^\mathsf{T}\mathsf{Y}}{\|\mathsf{R}_\varphi(\mathsf{Y})\|_V}=\frac{a(u_h,y_h)}{\|y_h\|_V}$, so that (50.11a) follows from the inf-sup and boundedness conditions on a. The proof of (50.11b) is similar since $(\mathcal{B}^\mathsf{T}\mathsf{P})^\mathsf{T}\mathsf{Y}=(\mathcal{B}\mathsf{Y})^\mathsf{T}\mathsf{P}=b(y_h,p_h)$ where $p_h=\mathsf{R}_\psi(\mathsf{P})$.
(ii) We observe that

$$\|\mathsf{R}_\varphi(\mathsf{U})\|_V^2=(\mathsf{R}_\varphi(\mathsf{U}),\mathsf{R}_\varphi(\mathsf{U}))_V=(\mathcal{J}_V\mathsf{U})^\mathsf{T}\mathsf{U}=(\mathcal{J}_V^{\frac12}\mathsf{U})^\mathsf{T}(\mathcal{J}_V^{\frac12}\mathsf{U})=\|\mathcal{J}_V^{\frac12}\mathsf{U}\|_{\ell^2(\mathbb{R}^N)}^2,$$

proving the first identity. The second identity results from

$$\begin{aligned}\|\mathcal{J}_V^{\frac12}\mathsf{U}\|_{\ell^2(\mathbb{R}^N)}&=\sup_{\mathsf{Y}\in\mathbb{R}^N}\frac{(\mathcal{J}_V^{\frac12}\mathsf{U})^\mathsf{T}\mathsf{Y}}{\|\mathsf{R}_\varphi(\mathsf{Y})\|_V}=\sup_{\mathsf{Y}\in\mathbb{R}^N}\frac{\mathsf{U}^\mathsf{T}\mathsf{Y}}{\|\mathsf{R}_\varphi(\mathcal{J}_V^{-\frac12}\mathsf{Y})\|_V}\\&=\sup_{\mathsf{Y}\in\mathbb{R}^N}\frac{\mathsf{U}^\mathsf{T}\mathsf{Y}}{\|\mathsf{Y}\|_{\ell^2(\mathbb{R}^N)}}=\|\mathsf{U}\|_{\ell^2(\mathbb{R}^N)}.\end{aligned}$$

Exercise 50.5 ($\mathcal{S}_\rho$). (i) A direct calculation shows that

$$\begin{aligned}(\mathcal{A}+\rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathcal{B})(\mathcal{A}^{-1}-\rho\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1})&=\mathcal{I}+\rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathcal{B}\mathcal{A}^{-1}\\&\quad-\rho\mathcal{B}^\mathsf{T}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1}-\rho^2\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathcal{S}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1}.\end{aligned}$$

The last term on the right-hand side is transformed as follows:

$$\begin{aligned}\rho^2\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathcal{S}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1}&=\rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}(-\mathcal{M}_Q+\mathcal{M}_Q+\rho\mathcal{S})(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1}\\&=-\rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathcal{B}\mathcal{A}^{-1}+\rho\mathcal{B}^\mathsf{T}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1}.\end{aligned}$$

Hence,

$$(\mathcal{A}+\rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathcal{B})(\mathcal{A}^{-1}-\rho\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1})=\mathcal{I}.$$

Similarly, one proves that $(\mathcal{A}^{-1}-\rho\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1})(\mathcal{A}+\rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathcal{B})=\mathcal{I}$.
(ii) Multiplying the above relation by $\mathcal{B}$ on the left and by $\mathcal{B}^\mathsf{T}$ on the right, we obtain $\mathcal{S}_\rho=\mathcal{S}-\rho\mathcal{S}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{S}$.
(iii) Following the hint, we obtain

$$\begin{aligned}&(\mathcal{S}^{-1}+\rho\mathcal{M}_Q^{-1})(\mathcal{S}-\rho\mathcal{S}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{S})\\&=\mathcal{I}+\rho\mathcal{M}_Q^{-1}\mathcal{S}-\rho(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{S}-\rho^2\mathcal{M}_Q^{-1}\mathcal{S}(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{S}\\&=\mathcal{I}+\rho\mathcal{M}_Q^{-1}\mathcal{S}-\rho\mathcal{M}_Q^{-1}(\mathcal{M}_Q+\rho\mathcal{S})(\mathcal{M}_Q+\rho\mathcal{S})^{-1}\mathcal{S}\\&=\mathcal{I}.\end{aligned}$$

The other identity is proved similarly.

Exercise 50.6 (Penalty). (i) Subtracting (50.18) from (50.9) yields

$$\begin{aligned}\mathcal{A}(\mathsf{U}-\mathsf{U}_\epsilon)+\mathcal{B}^\mathsf{T}(\mathsf{P}-\mathsf{P}_\epsilon)&=0,\\\mathcal{B}(\mathsf{U}-\mathsf{U}_\epsilon)+\epsilon\mathcal{M}_Q\mathsf{P}_\epsilon&=0.\end{aligned}$$

Using the inequalities (50.11a) and (50.11b) in the first equation yields

$$\|\mathsf{R}_\psi(\mathsf{P}-\mathsf{P}_\epsilon)\|_Q\le\frac{1}{\beta_h}\|\mathcal{B}^\mathsf{T}(\mathsf{P}-\mathsf{P}_\epsilon)\|_{\ell^2_\varphi}=\frac{1}{\beta_h}\|\mathcal{A}(\mathsf{U}-\mathsf{U}_\epsilon)\|_{\ell^2_\varphi}\le\frac{\|a\|}{\beta_h}\|\mathsf{R}_\varphi(\mathsf{U}-\mathsf{U}_\epsilon)\|_V.$$

Multiplying the first equation by $\mathsf{U}-\mathsf{U}_\epsilon$ and using the coercivity of a together with the second equation, we infer that

$$\begin{aligned}\alpha_h\,\|\mathsf{R}_\varphi(\mathsf{U}-\mathsf{U}_\epsilon)\|_V^2 &\le (\mathsf{U}-\mathsf{U}_\epsilon)^{\mathsf{T}}\mathcal{A}(\mathsf{U}-\mathsf{U}_\epsilon) = (\mathsf{U}-\mathsf{U}_\epsilon)^{\mathsf{T}}\mathcal{B}^{\mathsf{T}}(\mathsf{P}_\epsilon-\mathsf{P})\\ &= (\mathcal{B}(\mathsf{U}-\mathsf{U}_\epsilon))^{\mathsf{T}}(\mathsf{P}_\epsilon-\mathsf{P}) = -\epsilon(\mathcal{M}_Q\mathsf{P}_\epsilon)^{\mathsf{T}}(\mathsf{P}_\epsilon-\mathsf{P})\\ &= -\epsilon(\mathcal{M}_Q(\mathsf{P}_\epsilon-\mathsf{P}))^{\mathsf{T}}(\mathsf{P}_\epsilon-\mathsf{P}) - \epsilon(\mathcal{M}_Q\mathsf{P})^{\mathsf{T}}(\mathsf{P}_\epsilon-\mathsf{P})\\ &\le -\epsilon(\mathcal{M}_Q\mathsf{P})^{\mathsf{T}}(\mathsf{P}_\epsilon-\mathsf{P}) \le \epsilon\|\mathsf{R}_\psi(\mathsf{P}_\epsilon-\mathsf{P})\|_Q\|\mathsf{R}_\psi(\mathsf{P})\|_Q.\end{aligned}$$

Combining the above two inequalities yields (50.20).
(ii) The first estimate in the proof of Proposition 50.18 is unchanged:

$$\|\mathsf{R}_\psi(\mathsf{P}-\mathsf{P}_\epsilon)\|_Q \le \frac{\|a\|}{\beta_h}\|\mathsf{R}_\varphi(\mathsf{U}-\mathsf{U}_\epsilon)\|_V.$$

The second estimate becomes

$$\alpha_h\|\mathsf{R}_\varphi(\mathsf{U}-\mathsf{U}_\epsilon)\|_V^2 \le -\epsilon\lambda(\mathcal{I}_M\mathsf{P})^{\mathsf{T}}(\mathsf{P}_\epsilon-\mathsf{P}) \le \epsilon\lambda\|\mathsf{P}_\epsilon-\mathsf{P}\|_{\ell^2(\mathbb{R}^M)}\|\mathsf{P}\|_{\ell^2(\mathbb{R}^M)}.$$

Let $\mu_{\min}$ be the smallest eigenvalue of $\mathcal{M}_Q$. We have

$$\mu_{\min}\|\mathsf{Q}\|_{\ell^2(\mathbb{R}^M)}^2 \le \mathsf{Q}^{\mathsf{T}}\mathcal{M}_Q\mathsf{Q} = \|\mathsf{R}_\psi(\mathsf{Q})\|_Q^2.$$

This implies that

$$\alpha_h\|\mathsf{R}_\varphi(\mathsf{U}-\mathsf{U}_\epsilon)\|_V^2 \le \epsilon\lambda\mu_{\min}^{-1}\|\mathsf{R}_\psi(\mathsf{P}_\epsilon-\mathsf{P})\|_Q\|\mathsf{R}_\psi(\mathsf{P})\|_Q.$$

We infer that

$$\frac{\alpha_h\beta_h}{\|a\|}\|\mathsf{R}_\varphi(\mathsf{U}-\mathsf{U}_\epsilon)\|_V + \frac{\alpha_h\beta_h^2}{\|a\|^2}\|\mathsf{R}_\psi(\mathsf{P}-\mathsf{P}_\epsilon)\|_Q \le \epsilon\lambda\mu_{\min}^{-1}\|\mathsf{R}_\psi(\mathsf{P})\|_Q.$$

The method still converges when $\epsilon\to 0$, but to obtain a convergence rate close to ϵ, one should set $\lambda=\mu_{\min}$. If the mesh sequence is quasi-uniform, Proposition 28.6 shows that $\mu_{\min}\sim h^d$. Hence, one should choose $\lambda\sim h^d$. This method is interesting since it does not involve $\mathcal{M}_Q^{-1}$. More precisely, (50.19) can be rewritten as

$$\left(\mathcal{A}+\frac{1}{\lambda\epsilon}\mathcal{B}^{\mathsf{T}}\mathcal{B}\right)\mathsf{U}_\epsilon = \mathsf{F}+\frac{1}{\lambda\epsilon}\mathcal{B}^{\mathsf{T}}\mathsf{G}.$$

For instance, if the mass matrix $\mathcal{M}_Q$ can be lumped, which is the case for $\mathbb{P}_1$ and $\mathbb{Q}_1$ continuous finite elements, one could also use the lumped mass matrix instead of $h^d\mathcal{I}$.

Exercise 50.7 (Inexact Minres and DPG). (i) Let us set $\mathfrak{E}: V\to\mathbb{R}$ s.t. $\mathfrak{E}(v) := \frac{1}{2}\|f-B(v)\|_{Y'}^2$ for all $v\in V$. We have

$$\mathfrak{E}(v) = \frac{1}{2}\langle f-B(v), J_Y^{-1}(f-B(v))\rangle_{Y',Y}.$$

Since the sesquilinear form $a(v,w) := \langle B(v), J_Y^{-1}(B(w))\rangle_{Y',Y}$ is Hermitian and coercive, u minimizes $\mathfrak{E}$ over V iff $a(u,w) = \langle f, J_Y^{-1}(B(w))\rangle_{Y',Y}$ for all $w\in V$. Owing to the Lax–Milgram lemma, this problem admits a unique solution $u\in V$. Notice in passing that we have shown that $u\in V$ is the unique solution to the normal equation

$$B^*J_Y^{-1}B(u) = B^*J_Y^{-1}(f).$$

In what follows, we are going to construct an approximation of u.
(ii) Let us set $\mathfrak{E}_h : V_h \to \mathbb{R}$ s.t. $\mathfrak{E}_h(v_h) := \frac{1}{2}\|I_h^*(f - B(v_h))\|_{Y_h'}^2$. We have

$$\mathfrak{E}_h(v_h) = \frac{1}{2}\langle f - B(v_h), I_h J_{Y_h}^{-1} I_h^*(f - B(v_h))\rangle_{Y',Y},$$

so that $u_h \in V_h$ is characterized by the Euler equations

$$\langle f - B(u_h), I_h J_{Y_h}^{-1} I_h^* B(w_h)\rangle_{Y',Y} = 0, \qquad \forall w_h \in V_h.$$

Since $J_{Y_h}^{-1}$ is selfadjoint, these equations amount to

$$\langle B^*(I_h(r_h)), w_h\rangle_{V',V} = 0, \qquad \forall w_h \in V_h.$$

Moreover, the definition of r_h implies that for all $y_h \in Y_h$,

$$\begin{aligned}(r_h, y_h)_Y &= \langle J_{Y_h}(r_h), y_h\rangle_{Y_h',Y_h} = \langle f - B(u_h), I_h(y_h)\rangle_{Y',Y} \\ &= \langle I_h^* f, y_h\rangle_{Y_h',Y_h} - \langle I_h^* B(u_h), y_h\rangle_{Y_h',Y_h}.\end{aligned}$$

Thus, the pair $(u_h, r_h) \in V_h \times Y_h$ solves the following saddle point problem

$$\begin{aligned}&(r_h, y_h)_Y + \langle I_h^*(B(u_h)), y_h\rangle_{Y_h',Y_h} = \langle I_h^*(f), y_h\rangle_{Y_h',Y_h}, && \forall y_h \in Y_h, \\ &\langle B^*(I_h(r_h)), w_h\rangle_{V',V} = 0, && \forall w_h \in V_h.\end{aligned}$$

Since we have $\langle I_h^*(B(v_h)), y_h\rangle_{Y_h',Y_h} = b(v_h, y_h)$ for all $(v_h, y_h) \in V_h \times Y_h$, the inf-sup condition (50.1) implies that the above problem is well-posed.
(iii) Let us set $\mathfrak{G}_h : Y_h \to R$ s.t. $\mathfrak{G}_h(z_h) := \frac{1}{2}\|z_h\|_Y^2 - \langle I_h^*(f), z_h\rangle_{Y_h',Y_h}$. We have established in Proposition 49.11 that r_h minimizes $\mathfrak{G}_h$ in the subspace

$$(I_h^*(B(V_h)))^\perp = \{z_h \in Y_h \mid b(v_h, z_h) = 0,\ \forall v_h \in V_h\}$$

if and only if there is a unique Lagrange multiplier $u_h \in V_h$ such that the pair $(r_h, u_h) \in Y_h \times V_h$ is the unique solution to the above saddle point problem.
(iv) Recall that Lemma 26.9 shows that the inf-sup condition (50.1) implies the existence of a Fortin operator $\Pi_h : Y \to Y_h$ s.t. $b(v_h, y - \Pi_h(y)) = 0$ for all $v_h \in V_h$ and $\|\Pi_h\|_{\mathcal{L}(Y;Y_h)} \leq \frac{\|b\|}{\beta_0}$. We have

$$\beta\,\|u - u_h\|_V \leq \sup_{y \in Y} \frac{|b(u - u_h, y)|}{\|y\|_Y}.$$

Using that $B(u) = f$, we obtain

$$\begin{aligned}b(u - u_h, \Pi_h(y)) &= \langle B(u - u_h), I_h(\Pi_h(y))\rangle_{Y',Y} \\ &= \langle f - B(u_h), I_h(\Pi_h(y))\rangle_{Y',Y} \\ &= \langle I_h^*(f - B(u_h)), \Pi_h(y)\rangle_{Y_h',Y_h} \\ &= (J_{Y_h}^{-1}(I_h^*(f - B(u_h))), \Pi_h(y))_Y \\ &= (r_h, \Pi_h(y))_Y.\end{aligned}$$

Let w_h be arbitrary in U_h. We have

$$\begin{aligned}b(u - u_h, y) &= b(u - u_h, y - \Pi_h(y)) + b(u - u_h, \Pi_h(y)) \\ &= b(u - w_h, y - \Pi_h(y)) + (r_h, \Pi_h(y))_Y.\end{aligned}$$

This implies that

$$\beta \|u - u_h\|_V \leq \|b\|(1 + \|\Pi_h\|_{\mathcal{L}(Y;Y_h)})\|u - w_h\|_V + \|\Pi_h\|_{\mathcal{L}(Y;Y_h)}\|r_h\|_Y.$$

Moreover, we have

$$\|r_h\|_Y^2 = \langle I_h^* B(u - u_h), r_h\rangle_{Y_h',Y_h} = \langle I_h^* B(u - w_h), r_h\rangle_{Y_h',Y_h},$$

where we used the second equation of the saddle point problem. This shows that

$$\|r_h\|_Y \leq \|b\| \, \|u - w_h\|_V.$$

Putting everything together proves the quasi-optimal error estimate

$$\beta \|u - u_h\|_V \leq \|b\| \left(1 + 2\|\Pi_h\|_{\mathcal{L}(Y;Y_h)}\right) \inf_{w_h \in V_h} \|u - w_h\|_V.$$

Chapter 51

Darcy's equations

Exercises

Exercise 51.1 (Compactness). Let $D := (0,1)^3$ be the unit cube in $\mathbb{R}^3$. Show that the embedding $\boldsymbol{H}_0(\mathrm{div};D) \hookrightarrow \boldsymbol{L}^2(D)$ is not compact. (*Hint*: let

$$\begin{aligned}
\boldsymbol{\phi}_{1,n}(x_1,x_2,x_3) &:= \frac{1}{n\pi}\sin(n\pi x_2)\sin(n\pi x_3),\\
\boldsymbol{\phi}_{2,n}(x_1,x_2,x_3) &:= \frac{1}{n\pi}\sin(n\pi x_3)\sin(n\pi x_1),\\
\boldsymbol{\phi}_{3,n}(x_1,x_2,x_3) &:= \frac{1}{n\pi}\sin(n\pi x_1)\sin(n\pi x_2),
\end{aligned}$$

for all $n \geq 1$, set $\boldsymbol{v}_n := \nabla\times\boldsymbol{\phi}_n$, and prove first that $(\boldsymbol{v}_n)_{n\geq 1}$ weakly converges to zero in $\boldsymbol{L}^2(D)$ (see Definition C.28), then compute $\|\boldsymbol{v}_n\|_{\boldsymbol{L}^2(D)}$ and argue by contradiction.)

Exercise 51.2 (Neumann condition). Prove Proposition 51.3. (*Hint*: for the surjectivity of the divergence, solve a pure Neumann problem.)

Exercise 51.3 (Integration by parts). Let $H^1_{\mathrm{d}}(D)$ and $\boldsymbol{H}_{\mathrm{n}}(\mathrm{div};D)$ be defined in §51.1.3. Prove that $\int_D(\nabla q\cdot\boldsymbol{\varsigma} + q\nabla\cdot\boldsymbol{\varsigma})\,\mathrm{d}x = 0$ for all $q \in H^1_{\mathrm{d}}(D)$ and all $\boldsymbol{\varsigma} \in \boldsymbol{H}_{\mathrm{n}}(\mathrm{div};D)$. (*Hint*: observe that $\gamma^{\mathrm{g}}(q)_{|\partial D_{\mathrm{n}}} \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$.)

Exercise 51.4 (Primal, dual formulations). Prove Proposition 51.7.

Exercise 51.5 (Primal mixed formulation). Consider the problem: Find $p \in H^1(D)$ such that $-\Delta p = f$ and $\gamma^{\mathrm{g}}(p) = g$ with $f \in L^2(D)$ and $g \in H^{\frac{1}{2}}(\partial D)$. Derive a mixed formulation of this problem with unknowns $(p,\lambda) \in H^1(D)\times H^{-\frac{1}{2}}(\partial D)$ and show that it is well-posed. (*Hint*: set $b(v,\mu) := \langle\mu,\gamma^{\mathrm{g}}(v)\rangle_{\partial D}$ and observe that $B = \gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$.) Recover the PDE and the boundary condition. *Note*: this method is introduced in Babuška [2].

Exercise 51.6 (Fortin operator). Justify Remark 51.14. (*Hint*: use arguments similar to those of the proof of Lemma 51.10.)

Exercise 51.7 (Inf-sup condition). The goal is to prove the inf-sup condition (51.23) using the canonical Raviart–Thomas interpolation operator. (i) Do this by using elliptic regularity. (*Hint*: solve a Dirichlet problem.) (ii) Do this again by using the surjectivity of $\nabla\cdot : \boldsymbol{H}^1(D) \to L^2(D)$.

Exercise 51.8 (Error estimate). (i) Prove that

$$\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\boldsymbol{H}(\mathrm{div};D)} \le c_1' \inf_{\boldsymbol{\varsigma}_h\in\boldsymbol{V}_h} \|\boldsymbol{\sigma}-\boldsymbol{\varsigma}_h\|_{\boldsymbol{H}(\mathrm{div};D)},$$
$$\|p-p_h\|_{L^2(D)} \le c_3' \inf_{\boldsymbol{\varsigma}_h\in\boldsymbol{V}_h} \|\boldsymbol{\sigma}-\boldsymbol{\varsigma}_h\|_{\boldsymbol{H}(\mathrm{div};D)} + 2 \inf_{q_h\in Q_h} \|p-q_h\|_{L^2(D)},$$

with $c_1' := (1+\frac{\lambda_\sharp}{\lambda_\flat})(1+\frac{1}{\beta})$ and $c_3' := \frac{c_1'}{\lambda_\flat \beta_{L^2}'}$. (ii) Assuming that $\boldsymbol{\sigma}\in\boldsymbol{H}^r(D)$, $\nabla\cdot\boldsymbol{\sigma}\in H^r(D)$, and $p\in H^r(D)$ with $r\in(0,k+1]$, prove that

$$\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\boldsymbol{H}(\mathrm{div};D)} \le c\, h^r(|\boldsymbol{\sigma}|_{\boldsymbol{H}^r(D)} + |\nabla\cdot\boldsymbol{\sigma}|_{H^r(D)}),$$
$$\|p-p_h\|_{L^2(D)} \le c\, h^r(|\boldsymbol{\sigma}|_{\boldsymbol{H}^r(D)} + |\nabla\cdot\boldsymbol{\sigma}|_{H^r(D)} + |p|_{H^r(D)}).$$

(*Hint*: use the commuting projection $\mathcal{J}_h^{\mathrm{d}}$.)

Exercise 51.9 (Box scheme). Let $\mathbb{d} := \lambda_0\mathbb{I}_d$, $\lambda_0>0$, and enforce the boundary condition $\gamma^{\mathrm{g}}(p)=0$. Let $V_h := \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)\times P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$, where $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ is the Crouzeix–Raviart space defined in (36.8). Let $W_h := \boldsymbol{P}_0^{\mathrm{b}}(\mathcal{T}_h)\times P_0^{\mathrm{b}}(\mathcal{T}_h)$. Consider the bilinear form $a_h : V_h\times W_h\to\mathbb{R}$ defined by $a_h(v_h,w_h) := \lambda_0^{-1}(\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)} + (\nabla\cdot\boldsymbol{\sigma}_h,q_h)_{L^2(D)} + (\nabla_h p_h,\boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)}$ with $v_h := (\boldsymbol{\sigma}_h,p_h)$ and $w_h := (\boldsymbol{\tau}_h,q_h)$ (see Definition 36.3 for the broken gradient ∇_h). (i) Prove that $\dim(V_h)=\dim(W_h)$ and that there is $\alpha>0$ s.t. for all $v_h\in V_h$ and all $h\in\mathcal{H}$, $\alpha\|v_h\|_{V_h} \le \sup_{w_h\in W_h}\frac{|a_h(v_h,w_h)|}{\|w_h\|_{W_h}}$ with $\|v_h\|_{V_h}^2 := \lambda_0^{-1}\|\boldsymbol{\sigma}_h\|_{\boldsymbol{H}(\mathrm{div};D)}^2 + \lambda_0\|\nabla_h p_h\|_{\boldsymbol{L}^2(D)}^2$ and $\|w_h\|_{W_h}^2 := \lambda_0^{-1}\|\boldsymbol{\tau}_h\|_{\boldsymbol{L}^2(D)}^2 + \lambda_0\ell_D^{-2}\|q_h\|_{L^2(D)}^2$. (*Hint*: test with $(\underline{\boldsymbol{\sigma}}_h+\lambda_0\nabla_h p_h, 2\underline{p}_h+\ell_D^2\lambda_0^{-1}\nabla\cdot\boldsymbol{\sigma}_h)$, where $(\underline{\boldsymbol{\sigma}}_h,\underline{p}_h)$ is the L^2-orthogonal projection of $(\boldsymbol{\sigma}_h,p_h)$ onto W_h.) (ii) Consider the discrete problem: Find $u_h\in V_h$ such that $a_h(u_h,w_h) = (\boldsymbol{f},\boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)} + (g,q_h)_{L^2(D)}$ for all $w_h\in W_h$. Show that this problem is well-posed, prove a quasi-optimal error estimate, and show that the error converges to zero with rate h if the exact solution is smooth enough. (*Hint*: use Lemma 27.5.) *Note*: the scheme has been introduced in Croisille [12] to approximate (51.1). It is a Petrov–Galerkin scheme with only local test functions.

Solution to exercises

Exercise 51.1 (Compactness). Let $\boldsymbol{v}_n := \nabla\times\boldsymbol{\phi}_n$ with

$$\boldsymbol{\phi}_n(x_1,x_2,x_3) := \frac{1}{n\pi}\begin{pmatrix}\sin(n\pi x_2)\sin(n\pi x_3)\\ \sin(n\pi x_3)\sin(n\pi x_1)\\ \sin(n\pi x_1)\sin(n\pi x_2)\end{pmatrix},\quad \forall n\ge 1.$$

We obtain

$$\boldsymbol{v}_n = \begin{pmatrix}\sin(n\pi x_1)(\cos(n\pi x_2)-\cos(n\pi x_3))\\ \sin(n\pi x_2)(\cos(n\pi x_3)-\cos(n\pi x_1))\\ \sin(n\pi x_3)(\cos(n\pi x_1)-\cos(n\pi x_2))\end{pmatrix}.$$

We have $\boldsymbol{v}_n\in\boldsymbol{C}^\infty(D)$ and $\boldsymbol{v}_{n|\partial D}\cdot\boldsymbol{n}=\boldsymbol{0}$, so that $\boldsymbol{v}_n\in\boldsymbol{H}_0(\mathrm{div};D)$. Moreover, $\|\boldsymbol{v}_n\|_{\boldsymbol{L}^2(D)} = (\frac{3}{2})^{\frac12}$ and $\nabla\cdot\boldsymbol{v}_n=0$, so that $\|\boldsymbol{v}_n\|_{\boldsymbol{H}(\mathrm{div};D)} = (\frac{3}{2})^{\frac12}$. This means that the sequence $(\boldsymbol{v}_n)_{n\ge1}$ is bounded in $\boldsymbol{H}_0(\mathrm{div};D)$. Let us prove that the sequence $(\boldsymbol{v}_n)_{n\ge1}$ converges weakly to zero in $\boldsymbol{L}^2(D)$. For all $\boldsymbol{\phi}\in\boldsymbol{C}_0^\infty(D)$, we have

$$(\boldsymbol{v}_n,\boldsymbol{\phi})_{\boldsymbol{L}^2(D)} = -(\phi_n,\nabla\cdot\boldsymbol{\phi})_{L^2(D)} \to 0 \quad \text{as } n\to\infty.$$

Let now $\boldsymbol{w} \in \boldsymbol{L}^2(D)$. Owing to Theorem 1.38, for all $\epsilon > 0$, there is $\boldsymbol{\phi} \in \boldsymbol{L}^2(D)$ s.t.

$$\|\boldsymbol{w} - \boldsymbol{\phi}\|_{\boldsymbol{L}^2(D)} \le \epsilon$$

. Writing $(\boldsymbol{v}_n, \boldsymbol{w})_{\boldsymbol{L}^2(D)} = (\boldsymbol{v}_n, \boldsymbol{\phi})_{\boldsymbol{L}^2(D)} + (\boldsymbol{v}_n, \boldsymbol{w}-\boldsymbol{\phi})_{\boldsymbol{L}^2(D)}$ and using the Cauchy–Schwarz inequality to bound the second term, we infer that $\limsup_{n\to\infty} |(\boldsymbol{v}_n, \boldsymbol{w})_{\boldsymbol{L}^2(D)}| \le (\frac{3}{2})^{\frac{1}{2}}\epsilon$, and since $\epsilon > 0$ is arbitrary, we conclude that $\lim_{n\to\infty}(\boldsymbol{v}_n, \boldsymbol{w})_{\boldsymbol{L}^2(D)} = 0$. We have thus shown that the sequence $(\boldsymbol{v}_n)_{n\ge1}$ converges weakly to zero in $\boldsymbol{L}^2(D)$. We can now prove that the embedding $\boldsymbol{H}_0(\mathrm{div}; D) \hookrightarrow \boldsymbol{L}^2(D)$ is not compact. If the embedding were compact, there would exist a subsequence $(\boldsymbol{v}_{n_k})_{k\ge1}$ strongly converging to some $\boldsymbol{v} \in \boldsymbol{L}^2(D)$. But strong convergence implies weak convergence, so that $\boldsymbol{v} = \boldsymbol{0}$, and $\|\boldsymbol{v}_{n_k}\|_{\boldsymbol{L}^2(D)} = (\frac{3}{2})^{\frac{1}{2}}$ with strong convergence would also imply $\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} = (\frac{3}{2})^{\frac{1}{2}} > 0$, which is a contradiction.

Exercise 51.2 (Neumann condition). There are only two differences with the proof of Proposition 51.1. The first one concerns the divergence operator which now maps from $\boldsymbol{H}_0(\mathrm{div}; D)$ to $L^2_*(D)$ (owing to the divergence theorem). To prove that this operator is surjective, let $q \in L^2_*(D)$ and let us solve the pure Neumann problem $\phi \in H^1(D)$ such that $\Delta\phi = q$ and $\int_D \phi \,\mathrm{d}x = 0$. Then $\boldsymbol{\varsigma}_q = \nabla\phi$ has all the expected properties. The second difference concerns the boundary condition which is now a simple consequence of $\gamma^{\mathrm{d}}(\boldsymbol{\sigma}) = \gamma^{\mathrm{d}}(\boldsymbol{\sigma}_0) + \gamma^{\mathrm{d}}(\boldsymbol{\sigma}_{\mathrm{n}}) = a_{\mathrm{n}}$.

Exercise 51.3 (Integration by parts). Observe that

$$\int_D (\nabla q\cdot\boldsymbol{\varsigma} + q\nabla\cdot\boldsymbol{\varsigma})\,\mathrm{d}x = \langle\gamma^{\mathrm{d}}(\boldsymbol{\varsigma}), \gamma^{\mathrm{g}}(q)\rangle_{\partial D}.$$

We obtain $\gamma^{\mathrm{g}}(q)_{|\partial D_{\mathrm{n}}} \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$ since its zero-extension to ∂D is $\gamma^{\mathrm{g}}(q)$ (since $q \in H^1_{\mathrm{d}}(D)$) which is in $H^{\frac{1}{2}}(\partial D)$. This implies that

$$\langle\gamma^{\mathrm{d}}(\boldsymbol{\varsigma}), \gamma^{\mathrm{g}}(q)\rangle_{\partial D} = \langle\gamma^{\mathrm{d}}(\boldsymbol{\varsigma})_{|\partial D_{\mathrm{n}}}, \gamma^{\mathrm{g}}(q)_{|\partial D_{\mathrm{n}}}\rangle_{\partial D_{\mathrm{n}}},$$

and this last quantity vanishes since $\boldsymbol{\varsigma} \in \boldsymbol{H}_{\mathrm{n}}(\mathrm{div}; D)$.

Exercise 51.4 (Primal, dual formulations). Let $p \in H^1_0(D)$ solve (51.15) and define $\boldsymbol{\sigma} := \mathbb{d}^{-1}(\boldsymbol{f} - \nabla p)$. Since $\int_D(\nabla p\cdot\boldsymbol{\tau} + p\nabla\cdot\boldsymbol{\tau})\,\mathrm{d}x = 0$ for all $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}; D)$, the first equation of (51.6) is satisfied. Moreover, we infer that $\int_D \boldsymbol{\sigma}\cdot\nabla q\,\mathrm{d}x = -\int_D gq\,\mathrm{d}x$ for all $q \in H^1_0(D)$, implying that $\nabla\cdot\boldsymbol{\sigma} = g$ so that the second equation of (51.6) is satisfied. Let now $(\boldsymbol{\sigma}, p) \in \boldsymbol{H}(\mathrm{div}; D)\times L^2(D)$ solve (51.6). Then $\nabla\cdot\boldsymbol{\sigma} = g$ and taking a divergence-free test function $\boldsymbol{\tau}$ in the first equation of (51.6) shows that (51.17) is satisfied. Finally, let $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}; D)$ with $\nabla\cdot\boldsymbol{\sigma} = g$ solve (51.17), and let $p \in L^2(D)$ be s.t. $\int_D p\nabla\cdot\boldsymbol{\tau}\,\mathrm{d}x = \int_D(\mathbb{d}^{-1}\boldsymbol{\sigma} - \boldsymbol{f})\cdot\boldsymbol{\tau}\,\mathrm{d}x$ for all $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}; D)$ (note that the right-hand side vanishes if $\boldsymbol{\tau}$ is divergence-free). Since $\nabla\cdot : \boldsymbol{H}(\mathrm{div}; D) \to L^2(D)$ is surjective owing to Lemma 51.2, p is well defined, and its definition implies that the first equation of (51.6) is satisfied. The second one follows from $\nabla\cdot\boldsymbol{\sigma} = g$. Since $\nabla p = \boldsymbol{f} - \mathbb{d}^{-1}\boldsymbol{\sigma}$, the energy identity results from

$$\mathfrak{E}_\sharp(p) = -\frac{1}{2}\int_D \nabla p\cdot\mathbb{d}\cdot\nabla p\,\mathrm{d}x = -\frac{1}{2}\int_D(\boldsymbol{\varsigma} - \mathbb{d}\boldsymbol{f})\cdot\mathbb{d}^{-1}\cdot(\boldsymbol{\varsigma} - \mathbb{d}\boldsymbol{f}) = \mathfrak{E}_\flat(\boldsymbol{\sigma}).$$

Exercise 51.5 (Primal mixed formulation). The weak mixed formulation is

$$\begin{cases} \text{Find } p \in V := H^1(D) \text{ and } \lambda \in Q := H^{-\frac{1}{2}}(\partial D) \text{ such that} \\ a(p, q) + b(q, \lambda) = f(q), \qquad \forall q \in V, \\ b(p, \mu) = g(\mu), \qquad \forall \mu \in Q, \end{cases}$$

with $a(p,q) := \int_D \nabla p \cdot \nabla q \,\mathrm{d}x$, $b(v,\mu) := \langle \mu, \gamma^{\mathrm{g}}(v)\rangle_{\partial D}$, $f(q) := \int_D fq \,\mathrm{d}x$, and $g(\mu) := \langle \mu, g\rangle_{\partial D}$. All these forms are bounded. Moreover, the reflexivity of $H^{\frac{1}{2}}(\partial D)$ implies that $B = \gamma^{\mathrm{g}}$, so that $\ker(B) = \ker(\gamma^{\mathrm{g}}) = H^1_0(D)$ owing to Theorem 3.10. Hence, the bilinear form a is coercive on $\ker(B)$. Furthermore, B is surjective still by Theorem 3.10, so that the well-posedness follows from Theorem 49.13. The second equation in the weak mixed formulation implies that $\langle \mu, \gamma^{\mathrm{g}}(p) - g\rangle_{\partial D} = 0$ for all $\mu \in H^{-\frac{1}{2}}(\partial D)$, so that we recover the boundary condition $\gamma^{\mathrm{g}}(p) = g$ a.e. on ∂D, and the PDE follows by taking q arbitrary in $H^1_0(D)$ in the first equation.

Exercise 51.6 (Fortin operator). Let us verify the two properties stated in Lemma 26.9(i) with $W := \boldsymbol{H}(\mathrm{div}; D)$, $W_h := \boldsymbol{P}^{\mathrm{d}}_k(\mathcal{T}_h)$, $V := L^2(D)$, and $V_h := P^{\mathrm{b}}_k(\mathcal{T}_h)$. The operator Π_h maps from $\boldsymbol{H}(\mathrm{div}; D)$ to $\boldsymbol{P}^{\mathrm{d}}_k(\mathcal{T}_h)$ as required. Moreover, we infer that for all $q_h \in P^{\mathrm{b}}_k(\mathcal{T}_h)$ and all $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}; D)$,

$$\begin{aligned}\int_D q_h \nabla\cdot\Pi_h(\boldsymbol{v}) \,\mathrm{d}x &= \int_D q_h \nabla\cdot(\mathcal{J}^{\mathrm{d}}_h((\nabla\cdot)^\dagger(\mathcal{I}^{\mathrm{b}}_h(\nabla\cdot\boldsymbol{v})))) \,\mathrm{d}x\\ &= \int_D q_h \mathcal{J}^{\mathrm{b}}_h(\mathcal{I}^{\mathrm{b}}_h(\nabla\cdot\boldsymbol{v})) \,\mathrm{d}x\\ &= \int_D q_h \mathcal{I}^{\mathrm{b}}_h(\nabla\cdot\boldsymbol{v}) \,\mathrm{d}x\\ &= \int_D q_h \nabla\cdot\boldsymbol{v} \,\mathrm{d}x,\end{aligned}$$

since $\nabla\cdot\mathcal{J}^{\mathrm{d}}_h = \mathcal{J}^{\mathrm{b}}_h(\nabla\cdot)$, $P^{\mathrm{b}}_k(\mathcal{T}_h)$ is pointwise invariant under $\mathcal{J}^{\mathrm{b}}_h$, and q_h is in $P^{\mathrm{b}}_k(\mathcal{T}_h)$. In addition, using the stability of all the operators, one can see that $\|\Pi_h(\boldsymbol{v})\|_{\boldsymbol{H}(\mathrm{div};D)} \le c\|\nabla\cdot\boldsymbol{v}\|_{L^2(D)}$.

Exercise 51.7 (Inf-sup condition). (i) Let $q_h \in Q_h$. Let $\phi \in H^1_0(D)$ solve $\Delta\phi = q_h$ and set $\boldsymbol{\varsigma}_{q_h} := \nabla\phi$. Elliptic regularity implies that $\boldsymbol{\varsigma}_{q_h} \in \boldsymbol{H}^s(D)$ with $s > \frac{1}{2}$, i.e., $\boldsymbol{\varsigma}_{q_h} \in \check{\boldsymbol{V}}^{\mathrm{d}}(D)$, where the space $\check{\boldsymbol{V}}^{\mathrm{d}}(D)$ is defined in Lemma 19.6 (with $p := 2$). This means that $\boldsymbol{\varsigma}_{q_h}$ is in the domain of the interpolation operator $\mathcal{I}^{\mathrm{d}}_h$. Moreover, we have for all $K \in \mathcal{T}_h$,

$$\|\mathcal{I}^{\mathrm{d}}_K(\boldsymbol{\varsigma}_{q_h|K})\|_{\boldsymbol{L}^2(K)} \le c(\|\boldsymbol{\varsigma}_{q_h|K}\|_{\boldsymbol{L}^2(K)} + h_K^s|\boldsymbol{\varsigma}_{q_h|K}|_{\boldsymbol{H}^s(K)}).$$

Summing over the mesh cells and since h_K is bounded by the diameter of D, we infer that

$$\|\mathcal{I}^{\mathrm{d}}_h(\boldsymbol{\varsigma}_{q_h})\|_{\boldsymbol{L}^2(D)} \le c\|\boldsymbol{\varsigma}_{q_h}\|_{\boldsymbol{H}^s(D)} \le c'\|q\|_{L^2(D)}.$$

We can now conclude as in the proof of Lemma 51.10, but this time we take $\boldsymbol{\varsigma}^*_h = \mathcal{I}^{\mathrm{d}}_h(\boldsymbol{\varsigma}_{q_h})$.
(ii) Let again $q_h \in Q_h$. We use the hint to infer that there is $\boldsymbol{\varsigma}_{q_h} \in \boldsymbol{H}^1(D)$ such that $\nabla\cdot\boldsymbol{\varsigma}_{q_h} = q_h$ and $\|\boldsymbol{\varsigma}_{q_h}\|_{\boldsymbol{H}^1(D)} \le c\|q\|_{L^2(D)}$. Since

$$\|\mathcal{I}^{\mathrm{d}}_K(\boldsymbol{\varsigma}_{q_h|K})\|_{\boldsymbol{L}^2(K)} \le c(\|\boldsymbol{\varsigma}_{q_h|K}\|_{\boldsymbol{L}^2(K)} + h_K^1|\boldsymbol{\varsigma}_{q_h|K}|_{\boldsymbol{H}^1(K)})$$

for all $K \in \mathcal{T}_h$, we can now conclude as in Step (i).

Exercise 51.8 (Error estimate). (i) The error bound on $\boldsymbol{\sigma}$ follows from Corollary 50.5 since $\|a\| = \lambda_\sharp$, $\alpha = \lambda_\flat$, $\ker(B_h) \subset \ker(B)$, and we use Remark 50.6 to bound the norm of the Fortin operator by $\frac{\|b\|}{\beta} = \frac{1}{\beta}$. For the primal variable, we proceed as in the proof of Theorem 51.16 and use the bound on the dual variable.
(ii) To bound the best-approximation error on the dual variable, we use the commuting projection

$\mathcal{J}_h^{\mathrm{d}}$ and Theorem 23.12 to infer that

$$\begin{aligned}\|\boldsymbol{\sigma}-\mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\sigma})\|_{\boldsymbol{H}(\mathrm{div};D)} &\le \|\boldsymbol{\sigma}-\mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\sigma})\|_{\boldsymbol{L}^2(D)} + \|\nabla\cdot\boldsymbol{\sigma}-\mathcal{J}_h^{\mathrm{b}}(\nabla\cdot\boldsymbol{\sigma})\|_{L^2(D)}\\ &\le c\left(\inf_{\boldsymbol{\varsigma}_h\in\boldsymbol{P}_k^{\mathrm{d}}(\mathcal{T}_h)}\|\boldsymbol{\sigma}-\boldsymbol{\varsigma}_h\|_{\boldsymbol{L}^2(D)} + \inf_{\phi_h\in P_k^{\mathrm{b}}(\mathcal{T}_h)}\|\nabla\cdot\boldsymbol{\sigma}-\phi_h\|_{L^2(D)}\right),\end{aligned}$$

and we conclude invoking Corollary 22.9 (with $p:=2$ and $\mathrm{x}\in\{\mathrm{d},\mathrm{b}\}$).

Exercise 51.9 (Box scheme). (i) We observe that

$$\dim(V_h) = N_{\mathrm{f}} + N_{\mathrm{f}}^{\mathrm{i}} = (d+1)N_{\mathrm{c}} = \dim(W_h),$$

where N_{f} is the number of mesh faces, $N_{\mathrm{f}}^{\mathrm{i}}$ the number of mesh interfaces, and N_{c} the number of mesh cells. Let $v_h := (\boldsymbol{\sigma}_h, p_h)\in V_h$. Following the hint, we infer that

$$\begin{aligned}a_h(v_h,w_h) &= \lambda_0^{-1}(\boldsymbol{\sigma}_h, \underline{\boldsymbol{\sigma}}_h + \lambda_0\nabla_h p_h)_{\boldsymbol{L}^2(D)} + (\nabla\cdot\boldsymbol{\sigma}_h, 2\underline{p}_h + \ell_D^2\lambda_0^{-1}\nabla\cdot\boldsymbol{\sigma}_h)_{L^2(D)}\\ &\quad + (\nabla_h p_h, \underline{\boldsymbol{\sigma}}_h + \lambda_0\nabla_h p_h)_{\boldsymbol{L}^2(D)}\\ &= \lambda_0^{-1}(\boldsymbol{\sigma}_h,\underline{\boldsymbol{\sigma}}_h)_{\boldsymbol{L}^2(D)} + \ell_D^2\lambda_0^{-1}\|\nabla\cdot\boldsymbol{\sigma}_h\|_{L^2(D)}^2 + \lambda_0\|\nabla_h p_h\|_{\boldsymbol{L}^2(D)}^2\\ &\quad + 2(\nabla\cdot\boldsymbol{\sigma}_h,\underline{p}_h)_{L^2(D)} + (\boldsymbol{\sigma}_h,\nabla_h p_h)_{\boldsymbol{L}^2(D)} + (\nabla_h p_h,\underline{\boldsymbol{\sigma}}_h)_{\boldsymbol{L}^2(D)}\\ &= \lambda_0^{-1}(\|\underline{\boldsymbol{\sigma}}_h\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla\cdot\boldsymbol{\sigma}_h\|_{L^2(D)}^2) + \lambda_0\|\nabla_h p_h\|_{\boldsymbol{L}^2(D)}^2\\ &\quad + 2(\nabla\cdot\boldsymbol{\sigma}_h,p_h)_{L^2(D)} + 2(\boldsymbol{\sigma}_h,\nabla_h p_h)_{\boldsymbol{L}^2(D)}\\ &= \lambda_0^{-1}(\|\underline{\boldsymbol{\sigma}}_h\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla\cdot\boldsymbol{\sigma}_h\|_{L^2(D)}^2) + \lambda_0\|\nabla_h p_h\|_{\boldsymbol{L}^2(D)}^2,\end{aligned}$$

since $(\nabla\cdot\boldsymbol{\sigma}_h,\underline{p}_h)_{L^2(D)} = (\nabla\cdot\boldsymbol{\sigma}_h,p_h)_{L^2(D)}$, $(\underline{\boldsymbol{\sigma}}_h,\nabla_h p_h)_{\boldsymbol{L}^2(D)} = (\boldsymbol{\sigma}_h,\nabla_h p_h)_{\boldsymbol{L}^2(D)}$ and

$$(\nabla\cdot\boldsymbol{\sigma}_h,p_h)_{L^2(D)} + (\boldsymbol{\sigma}_h,\nabla_h p_h)_{\boldsymbol{L}^2(D)} = \sum_{F\in\mathcal{F}_h}(\boldsymbol{\sigma}_h\cdot\boldsymbol{n}_F,[\![p_h]\!])_{L^2(F)} = 0,$$

owing to the fact that the normal component of $\boldsymbol{\sigma}_h$ is continuous across F and takes a constant value on F and that $\int_F[\![p_h]\!]\,\mathrm{d}s = 0$ by definition of the space $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. Since $(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)_{|K} = \frac{1}{d}(\nabla\cdot\boldsymbol{\sigma}_h)(\boldsymbol{x}-\boldsymbol{x}_K)$, where $\boldsymbol{x}_K$ is the barycenter of K for all $K\in\mathcal{T}_h$, we infer that there is a uniform constant $c>0$ such that $\|\underline{\boldsymbol{\sigma}}_h\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla\cdot\boldsymbol{\sigma}_h\|_{L^2(D)}^2 \ge c(\|\boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla\cdot\boldsymbol{\sigma}_h\|_{L^2(D)}^2)$. This shows that

$$a_h(v_h,w_h)\ge\min(1,c)\,\|v_h\|_{V_h}^2,\qquad\forall v_h\in V_h.$$

Finally, we have

$$\|w_h\|_{W_h}^2 = \lambda_0^{-1}\|\underline{\boldsymbol{\sigma}}_h + \lambda_0\nabla_h p_h\|_{\boldsymbol{L}^2(D)}^2 + \lambda_0\ell_D^{-2}\|2\underline{p}_h + \ell_D^2\lambda_0^{-1}\nabla\cdot\boldsymbol{\sigma}_h\|_{L^2(D)}^2 \le c\,\|v_h\|_{V_h}^2,$$

where we used that $\|\underline{\boldsymbol{\sigma}}_h\|_{\boldsymbol{L}^2(D)}\le\|\boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}$, $\|\underline{p}_h\|_{L^2(D)}\le\|p_h\|_{\boldsymbol{L}^2(D)}$, and the discrete Poincaré–Steklov inequality $C_{\mathrm{PS}}^{\mathrm{CR}}\ell_D^{-1}\|p_h\|_{\boldsymbol{L}^2(D)}\le\|\nabla_h p_h\|_{\boldsymbol{L}^2(D)}$ (see Lemma 36.6). This proves the expected inf-sup condition on the bilinear form a_h. Notic e that $\|\cdot\|_{V_h}$ defines a norm on V_h owing to the discrete Poincaré–Steklov inequality.

(ii) Since the linear spaces V_h and W_h have the same dimension, the above inf-sup condition implies that the problem of finding $u_h\in V_h$ such that $a_h(u_h,w_h) = (\boldsymbol{f},\boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)} + (g,q_h)_{L^2(D)}$ for all $w_h\in W_h$, is well-posed. To establish an error estimate, we use Lemma 27.5. Let us set $V_\sharp := (\boldsymbol{H}(\mathrm{div};D)\times H_0^1(D)) + V_h$ which we equip with the norm

$$\|v\|_{V_\sharp}^2 := \lambda_0^{-1}\|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div};D)}^2 + \lambda_0\|\nabla_h p\|_{\boldsymbol{L}^2(D)}^2.$$

Notice that (27.5) holds true with $c_\sharp := 1$, i.e., $\|v_h\|_{V_\sharp} \le \|v_h\|_{V_h}$ for all $v_h \in V_h$. It remains to bound the consistency error. For all $(v_h, w_h) \in V_h \times W_h$, we have

$$\begin{aligned}\langle \delta_h(v_h), w_h\rangle_{W_h',W_h} &:= (\boldsymbol{f}, \boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)} + (g, q_h)_{L^2(D)} - a_h(v_h, w_h)\\ &= \lambda_0^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)} + (\nabla p - \nabla_h p_h, \boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)} + (\nabla\cdot\boldsymbol{\sigma} - \nabla\cdot\boldsymbol{\sigma}_h, q_h)_{L^2(D)}\\ &\le c\,\|(\boldsymbol{\sigma}, p) - (\boldsymbol{\sigma}_h, p_h)\|_{V_\sharp}\|w_h\|_{W_h},\end{aligned}$$

where we used the Cauchy–Schwarz inequality and the discrete Poincaré–Steklov inequality for the third term on the right-hand side. Owing to Lemma 27.5, we infer that there is a c s.t. for all $h \in \mathcal{H}$,

$$\|(\boldsymbol{\sigma}, p) - (\boldsymbol{\sigma}_h, p_h)\|_{V_\sharp} \le c \inf_{(\boldsymbol{\sigma}_h', p_h')\in V_h} \|(\boldsymbol{\sigma}, p) - (\boldsymbol{\sigma}_h', p_h')\|_{V_\sharp}.$$

If the solution is smooth enough, we can use the approximation properties of finite elements to infer that

$$\begin{aligned}\lambda_0^{-\frac12}(\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} + \ell_D\|\nabla\cdot(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{L^2(D)}) + \lambda_0^{\frac12}\|\nabla_h(p - p_h)\|_{\boldsymbol{L}^2(D)}&\\ \le c\,h\big(\lambda_0^{-\frac12}(|\boldsymbol{\sigma}|_{\boldsymbol{H}^1(D)} + \ell_D|\nabla\cdot\boldsymbol{\sigma}|_{H^1(D)}) + \lambda_0^{\frac12}|p|_{H^2(D)}\big),&\end{aligned}$$

that is, the error converges to zero with rate h.

Chapter 52

Potential and flux recovery

Exercises

Exercise 52.1 (Hybridization). Consider the discrete problem (52.4). (i) Let $\widetilde{Q}_h := Q_h \times \Lambda_h$ and $\widetilde{B}_h : \boldsymbol{V}_h^{\mathrm{hy}} \to \widetilde{Q}_h'$ s.t. $\langle \widetilde{B}_h(\boldsymbol{\tau}_h), (q_h, \mu_h)\rangle_{\widetilde{Q}_h', \widetilde{Q}_h} := b_h(\boldsymbol{\tau}_h, q_h) + c_h(\boldsymbol{\tau}_h, \mu_h)$ for all $\boldsymbol{\tau}_h \in \boldsymbol{V}_h^{\mathrm{hy}}$ and $(q_h, \mu_h) \in \widetilde{Q}_h$. Prove that $\widetilde{B}_h^*$ is injective. (*Hint*: integrate by parts and use the degrees of freedom of the $\mathbb{RT}_{k,d}$ element.) (ii) Prove that (52.4) admits a unique solution.

Exercise 52.2 (Crouzeix–Raviart). Assume that $\mathrm{d}_{|K}$ and $g_{|K}$ are constant over each mesh cell $K \in \mathcal{T}_h$. Let ∇_h denote the broken gradient (see Definition 36.3). Let $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ be the nonconforming Crouzeix–Raviart finite element space with homogeneous Dirichlet conditions (see (36.8)) and let $p_h^{\mathrm{CR}} \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ solve $\int_D (\mathrm{d}\nabla_h p_h^{\mathrm{CR}})\cdot\nabla_h q_h^{\mathrm{CR}}\,\mathrm{d}x = \int_D g q_h^{\mathrm{CR}}\,\mathrm{d}x$ for all $q_h^{\mathrm{CR}} \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. Let $\boldsymbol{x}_K$ be the barycenter of K for all $K \in \mathcal{T}_h$. Define

$$\begin{aligned}\boldsymbol{\sigma}_{h|K} &:= -(\mathrm{d}\nabla p_h^{\mathrm{CR}})_{|K} + d^{-1} g_{|K}(\boldsymbol{x} - \boldsymbol{x}_K)_{|K},\\ p_{h|K} &:= p_h^{\mathrm{CR}}(\boldsymbol{x}_K) + d^{-2}|K|^{-1} g_{|K}(\mathrm{d}^{-1}(\boldsymbol{x} - \boldsymbol{x}_K), \boldsymbol{x} - \boldsymbol{x}_K)_{L^2(K)}.\end{aligned}$$

(i) Prove that $\boldsymbol{\sigma}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$. (*Hint*: compute $\int_F [\![\boldsymbol{\sigma}_h]\!]\cdot\boldsymbol{n}_F \varphi_F^{\mathrm{CR}}\,\mathrm{d}s$ with φ_F^{CR} the Crouzeix–Raviart basis function attached to F.) (ii) Prove that $\int_D (q_h^{\mathrm{CR}} \nabla\cdot\boldsymbol{\tau}_h + \nabla_h q_h^{\mathrm{CR}}\cdot\boldsymbol{\tau}_h)\,\mathrm{d}x = 0$ for all $q_h^{\mathrm{CR}} \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ and all $\boldsymbol{\tau}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$. (iii) Prove that the pair $(\boldsymbol{\sigma}_h, p_h)$ solves (51.21) for $k := 0$ and $\boldsymbol{f} := \boldsymbol{0}$. (*Hint*: any function $\boldsymbol{\tau}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$ is such that $\boldsymbol{\tau}_{h|K} = \boldsymbol{\tau}_K + d^{-1}(\nabla\cdot\boldsymbol{\tau}_h)_{|K}(\boldsymbol{x} - \boldsymbol{x}_K)$, where $\boldsymbol{\tau}_K$ is the mean value of $\boldsymbol{\tau}_h$ on K.)

Exercise 52.3 (Post-processed potential). Let $k \geq 0$. Consider the simplicial Raviart–Thomas element $\mathbb{RT}_{k,d}$. Assume that it is possible to find a polynomial space $\mathbb{M}_{k,k'}$ so that for all $m \in \mathbb{M}_{k,k'}$, $\Pi_{Q_K}(m) = \Pi_{\Lambda_{\partial K}}(m) = 0$ implies that $m = 0$ for all $K \in \mathcal{T}_h$. Prove that $(\nabla m, \boldsymbol{\tau})_{\boldsymbol{L}^2(K)} = 0$ for all $\boldsymbol{\tau} \in \mathbb{RT}_{k,d}$ implies that $m = 0$. (*Hint*: integrate by parts and use the degrees of freedom in $\mathbb{RT}_{k,d}$.) Let now m_h^{nc} be the post-processed potential from the dual mixed formulation (52.2). Show that $\|\nabla m_h^{\mathrm{nc}}\|_{\boldsymbol{L}^2(K)} \leq c\|\mathrm{d}^{-1}\boldsymbol{\sigma}_h - \boldsymbol{f}\|_{\boldsymbol{L}^2(K)}$ for all $K \in \mathcal{T}_h$. (*Hint*: use norm equivalence on the reference element, then (52.13); see also Vohralík [45, Lem. 5.4].)

Exercise 52.4 (Bound (52.19)). Prove (52.19). (*Hint*: use Theorem 34.19.)

Exercise 52.5 (Inverse inequality). Prove (52.20). (*Hint*: consider the dual mixed formulation of (52.17) and introduce the post-processed variable $m_{\boldsymbol{z}}^{\mathrm{nc}}$, use (52.13), accept as a fact that $\|m_{\boldsymbol{z}}^{\mathrm{nc}}\|_{L^2(D_{\boldsymbol{z}})} \leq c h_{D_{\boldsymbol{z}}} \|\nabla_h m_{\boldsymbol{z}}^{\mathrm{nc}}\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}$, and bound traces of $m_{\boldsymbol{z}}^{\mathrm{nc}}$ using Lemma 12.15.)

Exercise 52.6 (Prager–Synge equality). Let $u \in H^1_0(D)$ be such that $-\Delta u = f$ in $L^2(D)$. Let $u_h \in H^1_0(D)$, and let $\boldsymbol{\sigma}^* \in \boldsymbol{H}(\mathrm{div}; D)$ be such that $\nabla\cdot\boldsymbol{\sigma}^* = f$. Prove that $\|\nabla(u-u_h)\|^2_{\boldsymbol{L}^2(D)} + \|\nabla u + \boldsymbol{\sigma}^*\|^2_{\boldsymbol{L}^2(D)} = \|\nabla u_h + \boldsymbol{\sigma}^*\|^2_{\boldsymbol{L}^2(D)}$. (*Hint*: compute $(\nabla(u-u_h), \nabla u + \boldsymbol{\sigma}^*)_{\boldsymbol{L}^2(D)}$.)

Solution to exercises

Exercise 52.1 (Hybridization). (i) Let $(q_h, \mu_h) \in \ker(\widetilde{B}_h^*)$, i.e., $b_h(\boldsymbol{\tau}_h, q_h) + c_h(\boldsymbol{\tau}_h, \mu_h) = 0$ for all $\boldsymbol{\tau}_h \in \boldsymbol{V}_h^{\mathrm{hy}}$. Integrating by parts in each mesh cell $K \in \mathcal{T}_h$, we infer that

$$\sum_{K\in\mathcal{T}_h} \int_K \nabla q_{h|K}\cdot\boldsymbol{\tau}_h \,\mathrm{d}x + \int_{\partial K} (\mu_{h|\partial K} - q_{h|K})(\boldsymbol{\tau}_h\cdot\boldsymbol{n}_K)\,\mathrm{d}s = 0.$$

Since $\nabla q_{h|K} \in \mathbb{P}_{k-1,d}$ for all $K \in \mathcal{T}_h$ and $(\mu_{h|F} - q_{h|K}) \circ \boldsymbol{T}_F \in \mathbb{P}_{k,d-1}$ for all $F \in \mathcal{F}_K$, we can use the degrees of freedom of the $\mathbb{RT}_{k,d}$ element and choose $\boldsymbol{\tau}_h \in \boldsymbol{V}_h^{\mathrm{hy}}$ to obtain

$$\sum_{K\in\mathcal{T}_h} \Big(\|\nabla q_h\|^2_{\boldsymbol{L}^2(K)} + h_K^{-1}\|\mu_{h|\partial K} - q_{h|K}\|^2_{L^2(\partial K)} \Big) = 0.$$

This implies that q_h is piecewise constant and that $\mu_{h|\partial K} = q_{h|K}$ on the boundary of each mesh cell. Since μ_h vanishes on the boundary faces, we infer that q_h vanishes on all the mesh cells having a boundary face and that μ_h vanishes on all the faces of those cells. We can repeat the argument for the cells sharing an interface with those cells, and we can move inward and reach all the cells in $\mathcal{T}_h$ by repeating this process a finite number of times. This proves that $q_h = 0$ and $\mu_h = 0$. Hence, $\widetilde{B}_h^*$ is injective.
(ii) The discrete problem (52.4) is a finite-dimensional saddle point problem. The bilinear form a is coercive on $\boldsymbol{V}_h^{\mathrm{hy}}\times\boldsymbol{V}_h^{\mathrm{hy}}$ and the bilinear form $\widetilde{b}_h$ associated with the operator $\widetilde{B}_h$ on $\boldsymbol{V}_h^{\mathrm{hy}}\times\widetilde{Q}_h$ (i.e., $\widetilde{b}_h(\boldsymbol{\tau}_h, (q_h,\mu_h)) := \langle \widetilde{B}_h(\boldsymbol{\tau}_h), (q_h,\mu_h)\rangle_{\widetilde{Q}_h', \widetilde{Q}_h}$) satisfies a discrete inf-sup condition owing to the injectivity of the adjoint operator $\widetilde{B}_h^*$ established in Step (i). Hence, the discrete problem (52.4) admits a unique solution.

Exercise 52.2 (Crouzeix–Raviart). (i) Let $F \in \mathcal{F}_h^\circ$ and let D_F be composed of the points in the two cells such that $F := \partial K_l \cap \partial K_r$. Observe from the definition of $\boldsymbol{\sigma}_h$ that $\boldsymbol{\sigma}_h\cdot\boldsymbol{n}_{F'}$ is piecewise constant on each face $F' \subset \partial D_F$ and that $\nabla\cdot\boldsymbol{\sigma}_{h|K} = g_{|K}$. Following the hint, we infer that

$$\begin{aligned}
\int_F [\![\boldsymbol{\sigma}_h]\!]\cdot\boldsymbol{n}_F \varphi_F^{\mathrm{CR}}\,\mathrm{d}s &= \int_{K_l\cup K_r} \nabla\cdot(\boldsymbol{\sigma}_h\varphi_F^{\mathrm{CR}})\,\mathrm{d}x \\
&= \int_{K_l\cup K_r} ((\nabla\cdot\boldsymbol{\sigma}_h)\varphi_F^{\mathrm{CR}} + \boldsymbol{\sigma}_h\cdot\nabla\varphi_F^{\mathrm{CR}})\,\mathrm{d}x \\
&= \int_{K_l\cup K_r} (g\varphi_F^{\mathrm{CR}} - \nabla p_h^{\mathrm{CR}}\cdot\mathrm{d}\cdot\nabla\varphi_F^{\mathrm{CR}})\,\mathrm{d}x = 0,
\end{aligned}$$

since $\int_{F'} \boldsymbol{\sigma}_h\cdot\boldsymbol{n}_F\varphi_F^{\mathrm{CR}}\,\mathrm{d}s = 0$ (because φ_F^{CR} has zero mean value on each face $F' \subset \partial D_F$) and $\int_{K'}(\boldsymbol{x} - \boldsymbol{x}_{K'})\cdot\nabla\varphi_h^{\mathrm{CR}}\,\mathrm{d}x = 0$ (because $\nabla\varphi_h^{\mathrm{CR}}$ is piecewise constant). Observing that both $[\![\boldsymbol{\sigma}_h]\!]\cdot\boldsymbol{n}_F$ and φ_F^{CR} are constant on F, we infer that $[\![\boldsymbol{\sigma}_h]\!]\cdot\boldsymbol{n}_F = 0$. Since, by its definition, $\boldsymbol{\sigma}_h \in \boldsymbol{P}_0^{\mathrm{d,b}}(\mathcal{T}_h)$, we conclude that $\boldsymbol{\sigma}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$.

(ii) We integrate by parts in each cell $K \in \mathcal{T}_h$. This yields

$$\int_D (q_h^{\mathrm{CR}} \nabla\cdot\boldsymbol{\tau}_h + \nabla_h q_h^{\mathrm{CR}} \cdot \boldsymbol{\tau}_h)\,\mathrm{d}x = \sum_{F\in\mathcal{F}_h^\circ} \int_F \boldsymbol{\tau}_h\cdot\boldsymbol{n}_F [\![q_h^{\mathrm{CR}}]\!]_F \,\mathrm{d}s + \sum_{F\in\mathcal{F}_h^\partial} \int_F \boldsymbol{\tau}_h\cdot\boldsymbol{n}_F q_h^{\mathrm{CR}}\,\mathrm{d}s,$$

since the normal component of $\boldsymbol{\tau}_h$ is single-valued at interfaces. To conclude that the right-hand side is zero, we observe that the normal component of $\boldsymbol{\tau}_h$ is constant on all the faces, whereas $[\![q_h^{\mathrm{CR}}]\!]_F$ and $q_{h|F}^{\mathrm{CR}}$ have zero mean value on all $F \in \mathcal{F}_h^\circ$ and $F \in \mathcal{F}_h^\partial$, respectively.
(iii) Since $\boldsymbol{\sigma}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$ satisfies $\nabla\cdot\boldsymbol{\sigma}_h = g$, it remains to show that the weak form of Darcy's law is satisfied. Let $\boldsymbol{\tau}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$. We infer that

$$\begin{aligned}
\int_K p_h \nabla\cdot\boldsymbol{\tau}_h\,\mathrm{d}x &= \int_K p_h^{\mathrm{CR}}(\boldsymbol{x}_K)\nabla\cdot\boldsymbol{\tau}_h\,\mathrm{d}x + \int_K d^{-2} g_{|K}(\boldsymbol{x}-\boldsymbol{x}_K)\cdot(\mathbb{d}^{-1}(\boldsymbol{x}-\boldsymbol{x}_K))\nabla\cdot\boldsymbol{\tau}_h\,\mathrm{d}x\\
&= \int_K p_h^{\mathrm{CR}}\nabla\cdot\boldsymbol{\tau}_h\,\mathrm{d}x + \int_K d^{-1} g_{|K}(\boldsymbol{x}-\boldsymbol{x}_K)\cdot(\mathbb{d}^{-1}(\boldsymbol{\tau}_h-\boldsymbol{\tau}_K))\,\mathrm{d}x\\
&= \int_K p_h^{\mathrm{CR}}\nabla\cdot\boldsymbol{\tau}_h\,\mathrm{d}x + \int_K d^{-1} g_{|K}(\boldsymbol{x}-\boldsymbol{x}_K)\cdot(\mathbb{d}^{-1}\boldsymbol{\tau}_h)\,\mathrm{d}x\\
&= \int_K p_h^{\mathrm{CR}}\nabla\cdot\boldsymbol{\tau}_h\,\mathrm{d}x + \int_K (\boldsymbol{\sigma}_h + \mathbb{d}\nabla p_h^{\mathrm{CR}})\cdot(\mathbb{d}^{-1}\boldsymbol{\tau}_h)\,\mathrm{d}x,
\end{aligned}$$

where we used the definition of p_h and the fact that g and $\nabla\cdot\boldsymbol{\tau}_h$ are constant on K in the first line, that $\int_K p_h^{\mathrm{CR}}\,\mathrm{d}x = p_h^{\mathrm{CR}}(\boldsymbol{x}_K)$, $\nabla\cdot\boldsymbol{\tau}_h$ is constant on K, and the hint in the second line, and that g is constant on K and the definition of $\boldsymbol{\sigma}_h$ in the third line. Summing over $K \in \mathcal{T}_h$ and using Step (ii), we infer that $\int_D p_h\nabla\cdot\boldsymbol{\tau}_h\,\mathrm{d}x = \int_K \boldsymbol{\tau}_h\cdot(\mathbb{d}^{-1}\boldsymbol{\sigma}_h)\,\mathrm{d}x$.

Exercise 52.3 (Post-processed potential). Let $m \in \mathbb{M}_{k,k'}$ and let us assume that $(\nabla m, \boldsymbol{\tau})_{\boldsymbol{L}^2(K)} = 0$ for all $\boldsymbol{\tau} \in \mathbb{RT}_{k,d}$. Integrating by parts, we infer that

$$\begin{aligned}
0 = (\nabla m, \boldsymbol{\tau})_{\boldsymbol{L}^2(K)} &= -(m, \nabla\cdot\boldsymbol{\tau})_{L^2(K)} + (m, \boldsymbol{\tau}\cdot\boldsymbol{n}_K)_{L^2(\partial K)}\\
&= -(\Pi_{Q_K}(m), \nabla\cdot\boldsymbol{\tau})_{L^2(K)} + (\Pi_{\Lambda_{\partial K}}(m), \boldsymbol{\tau}\cdot\boldsymbol{n}_K)_{L^2(\partial K)},
\end{aligned}$$

since $\nabla\cdot\boldsymbol{\tau} \in \mathbb{P}_{k,d} = Q_K$ and $\boldsymbol{\tau}\cdot\boldsymbol{n}_K$ is a piecewise polynomial of degree at most k on the faces in ∂K. Subtracting the mean value $\underline{m}_K := \frac{1}{|K|}\int_K m\,\mathrm{d}x$ and using the divergence theorem, we obtain with $m' := m - \underline{m}_K$,

$$0 = -(\Pi_{Q_K}(m'), \nabla\cdot\boldsymbol{\tau})_{L^2(K)} + (\Pi_{\Lambda_{\partial K}}(m'), \boldsymbol{\tau}\cdot\boldsymbol{n}_K)_{L^2(\partial K)}.$$

Let us consider a function $\boldsymbol{\tau} \in \mathbb{RT}_{k,d}$ such that $\boldsymbol{\tau}\cdot\boldsymbol{n}_K = 0$. Integrating by parts again, we infer that $\nabla\Pi_{Q_K}(m') = \boldsymbol{0}$ since this function is in $\mathbb{P}_{k-1,d}$ and moments in K against functions in $\mathbb{P}_{k-1,d}$ are possible the degrees of freedom in $\mathbb{RT}_{k,d}$ once those attached to faces have been set to zero. Since m' has zero-mean value in K and Π_{Q_K} preserves this property, we infer that $\Pi_{Q_K}(m') = 0$. We now obtain that $0 = (\Pi_{\Lambda_{\partial K}}(m'), \boldsymbol{\tau}\cdot\boldsymbol{n}_K)_{L^2(\partial K)}$, and choosing now $\boldsymbol{\tau}$ to have arbitrary normal component on each of the faces in ∂K, we infer that $\Pi_{\Lambda_{\partial K}}(m') = 0$. We can now use the assumption on Π_{Q_K} and $\Pi_{\Lambda_{\partial K}}$ to infer that $m = \underline{m}_K$, i.e., m is constant.
Mapping to the reference element and using norm equivalence, we infer that there is a constant c s.t. for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$,

$$\|\nabla m\|_{\boldsymbol{L}^2(K)} \le c \sup_{\boldsymbol{\tau}\in\mathbb{RT}_{k,d}} \frac{|(\nabla m, \boldsymbol{\tau})_{\boldsymbol{L}^2(K)}|}{\|\boldsymbol{\tau}\|_{\boldsymbol{L}^2(K)}}.$$

Let now m_h^{nc} be the post-processed potential from the dual mixed formulation (52.2). Using (52.13), we infer that

$$\begin{aligned}\|\nabla m_h^{\mathrm{nc}}\|_{\boldsymbol{L}^2(K)} &\le c \sup_{\boldsymbol{\tau}\in\mathbf{RT}_{k,d}} \frac{|(\nabla m_h^{\mathrm{nc}},\boldsymbol{\tau})_{\boldsymbol{L}^2(K)}|}{\|\boldsymbol{\tau}\|_{\boldsymbol{L}^2(K)}}\\ &= c \sup_{\boldsymbol{\tau}\in\mathbf{RT}_{k,d}} \frac{|(\mathbb{d}^{-1}\boldsymbol{\sigma}_h-\boldsymbol{f},\boldsymbol{\tau})_{\boldsymbol{L}^2(K)}|}{\|\boldsymbol{\tau}\|_{\boldsymbol{L}^2(K)}}\\ &\le \|\mathbb{d}^{-1}\boldsymbol{\sigma}_h-\boldsymbol{f}\|_{\boldsymbol{L}^2(K)}.\end{aligned}$$

Exercise 52.4 (Bound (52.19)**).** Recalling the notation in Corollary 34.14, the key point is that, upon defining

$$\begin{aligned}\eta_{K'}^{\mathrm{v}}(p_h) &:= h_{K'}\|g+\nabla\cdot(\mathbb{d}\nabla p_h)\|_{L^2(K')},\\ \tilde\eta_{K'}^{\mathrm{s}}(p_h) &:= \Big(\sum_{F'\in\mathcal{F}_{K'}\cap\check{\mathcal{F}}_K^\circ} h_{F'}\|[\![\mathbb{d}\nabla p_h]\!]\|_{L^2(F')}^2\Big)^{\frac12},\end{aligned}$$

the bound (52.18) can be rewritten as follows, where $c>0$ only depends on the regularity of the mesh sequence:

$$c\,\|\boldsymbol{\sigma}_h^*+\mathbb{d}\nabla p_h\|_{\boldsymbol{L}^2(K)} \le \sum_{K'\in\check{\mathcal{T}}_K}(\eta_{K'}^{\mathrm{v}}(p_h)+\tilde\eta_{K'}^{\mathrm{s}}(p_h)).$$

We can now use Theorem 34.19 to bound the right-hand side: the oscillation terms from Definition 34.17 reduce to $\omega_{K'}^{\mathrm{v}} = h_{K'}\|g-\mathcal{I}_l^{\mathrm{b}}(g)\|_{L^2(K')}$ (with $l^{\mathrm{v}}:=l\ge k-1$) and $\omega_{K'}^{\mathrm{s}}=0$ (with $l^{\mathrm{s}}:=k-1$) since $\mathbb{d}$ is piecewise constant, and only the cells in $\mathcal{T}_{\boldsymbol{z}}$ need to be considered when bounding $\tilde\eta_{K'}^{\mathrm{s}}(p_h)$ since only jumps across interfaces in $\check{\mathcal{F}}_K^\circ$ are involved.

Exercise 52.5 (Inverse inequality). Following the hint, let $m_{\boldsymbol{z}}^{\mathrm{nc}}$ be the post-processed variable from the dual mixed formulation of (52.17). If $\boldsymbol{z}\in\mathcal{V}_h^\circ$, $m_{\boldsymbol{z}}^{\mathrm{nc}}$ has zero mean value in $D_{\boldsymbol{z}}$, and both $\boldsymbol{\sigma}_{\boldsymbol{z}}^*$ and $\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})$ have zero normal component on $\partial D_{\boldsymbol{z}}$ (recall that $\boldsymbol{f}_{\boldsymbol{z}}=-\psi_{\boldsymbol{z}}\mathbb{d}\nabla p_h$, that $\psi_{\boldsymbol{z}}$ vanishes at $\partial D_{\boldsymbol{z}}$, and that $\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}$ preserves zero normal components). If $\boldsymbol{z}\in\mathcal{V}_h^\partial$, $m_{\boldsymbol{z}}^{\mathrm{nc}}$ has zero moments up to order l on all faces located in $\partial D_{\boldsymbol{z}}\cap\partial D$, and both $\boldsymbol{\sigma}_{\boldsymbol{z}}^*$ and $\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})$ have zero normal component on all faces located in $\partial D_{\boldsymbol{z}}\setminus\partial D$. Owing to (52.13), we infer that

$$\|\boldsymbol{\sigma}_{\boldsymbol{z}}^*-\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})\|_{L^2(D_{\boldsymbol{z}})}^2 = (\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})-\boldsymbol{\sigma}_{\boldsymbol{z}}^*,\nabla_h m_{\boldsymbol{z}}^{\mathrm{nc}})_{L^2(D_{\boldsymbol{z}})}.$$

Integrating by parts and using the above properties on $\partial D_{\boldsymbol{z}}$ as well as

$$\nabla\cdot(\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})-\boldsymbol{\sigma}_{\boldsymbol{z}}^*) = \mathcal{I}_l^{\mathrm{b}}(\nabla\cdot\boldsymbol{f}_{\boldsymbol{z}}-g_{\boldsymbol{z}})$$

on each cell in $D_{\boldsymbol{z}}$, we infer that

$$\begin{aligned}\|\boldsymbol{\sigma}_{\boldsymbol{z}}^*-\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})\|_{L^2(D_{\boldsymbol{z}})}^2 &= -(\mathcal{I}_l^{\mathrm{b}}(\nabla\cdot\boldsymbol{f}_{\boldsymbol{z}}-g_{\boldsymbol{z}}),m_{\boldsymbol{z}}^{\mathrm{nc}})_{L^2(D_{\boldsymbol{z}})}\\ &\quad+\sum_{F'\in\mathcal{F}_{\boldsymbol{z}}^\circ}\int_{F'}[\![m_h^{\mathrm{nc}}(\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})-\boldsymbol{\sigma}_{\boldsymbol{z}}^*)]\!]\cdot\boldsymbol{n}_{F'}\,\mathrm{d}s.\end{aligned}$$

Recalling that the jump of m_h^{nc} has zero moments up to order l, that the normal component of $\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})-\boldsymbol{\sigma}_{\boldsymbol{z}}^*$ in each cell in $D_{\boldsymbol{z}}$ is a polynomial of order at most l, and that $[\![\boldsymbol{\sigma}_{\boldsymbol{z}}^*]\!]\cdot\boldsymbol{n}_{F'}=0$, we infer that

$$\int_{F'}[\![m_h^{\mathrm{nc}}(\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})-\boldsymbol{\sigma}_{\boldsymbol{z}}^*)]\!]\cdot\boldsymbol{n}_{F'}\,\mathrm{d}s = (\{m_h^{\mathrm{nc}}\},[\![\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})]\!]\cdot\boldsymbol{n}_{F'})_{L^2(F')}.$$

We now bound $(\mathcal{I}_l^{\mathrm{b}}(\nabla\cdot\boldsymbol{f}_{\boldsymbol{z}}-g_{\boldsymbol{z}}), m_{\boldsymbol{z}}^{\mathrm{nc}})_{L^2(D_{\boldsymbol{z}})}$ and $(\{m_h^{\mathrm{nc}}\}, [\![\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})]\!]\cdot\boldsymbol{n}_{F'})_{L^2(F')}$ using Cauchy–Schwarz inequalities. We use the inequality $\|m_{\boldsymbol{z}}^{\mathrm{nc}}\|_{L^2(D_{\boldsymbol{z}})} \le c h_{D_{\boldsymbol{z}}} \|\nabla_h m_{\boldsymbol{z}}^{\mathrm{nc}}\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}$ given in the hint (this broken Poincaré–Steklov inequality can be proven along the lines of Exercise 22.3), and we combine it with the multiplicative trace inequality from Lemma 12.15 and the regularity of the mesh sequence to infer that

$$\|m_{\boldsymbol{z}}^{\mathrm{nc}}\|_{L^2(F')} \le c\, h_{F'}^{\frac{1}{2}} \|\nabla_h m_{\boldsymbol{z}}^{\mathrm{nc}}\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}.$$

Using the stability result from Exercise 52.3, we can bound $\|\nabla_h m_{\boldsymbol{z}}^{\mathrm{nc}}\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}$ by $\|\boldsymbol{\sigma}_{\boldsymbol{z}}^* - \boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})\|_{L^2(D_{\boldsymbol{z}})}$. This leads to

$$c\,\|\boldsymbol{\sigma}_{\boldsymbol{z}}^* - \boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})\|_{L^2(D_{\boldsymbol{z}})} \le h_{D_{\boldsymbol{z}}} \|\mathcal{I}_l^{\mathrm{b}}(\nabla\cdot\boldsymbol{f}_{\boldsymbol{z}} - g_{\boldsymbol{z}})\|_{L^2(D_{\boldsymbol{z}})} + \sum_{F'\in\mathcal{F}_{\boldsymbol{z}}^{\circ}} h_{F'}^{\frac{1}{2}} \|[\![\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})]\!]\cdot\boldsymbol{n}_{F'}\|_{L^2(F')},$$

with $c>0$, and we conclude using the L^2-stability of $\mathcal{I}_l^{\mathrm{b}}$, the fact that the normal components of $\boldsymbol{\mathcal{I}}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})$ are L^2-orthogonal projections of those of $\boldsymbol{f}_{\boldsymbol{z}}$, and the regularity of the mesh sequence.

Exercise 52.6 (Prager–Synge equality). $(\nabla(u-u_h), \nabla u + \boldsymbol{\sigma}^*)_{\boldsymbol{L}^2(D)} = 0$ follows from integration by parts since $(u-u_h) \in H_0^1(D)$ and $\nabla\cdot(\nabla u + \boldsymbol{\sigma}^*) = -f + f = 0$. As a result, we have

$$\begin{aligned}\|\nabla u_h + \boldsymbol{\sigma}^*\|_{\boldsymbol{L}^2(D)}^2 &= \|\nabla(u-u_h) - (\nabla u + \boldsymbol{\sigma}^*)\|_{\boldsymbol{L}^2(D)}^2 \\ &= \|\nabla(u-u_h)\|_{\boldsymbol{L}^2(D)}^2 + \|\nabla u + \boldsymbol{\sigma}^*\|_{\boldsymbol{L}^2(D)}^2.\end{aligned}$$

This proves the assertion.

Chapter 53

Stokes equations: Basic ideas

Exercises

Exercise 53.1 ($\nabla\cdot$ is surjective). Let $D \subset \mathbb{R}^2$ be a domain of class C^2. Prove that $\nabla\cdot : \boldsymbol{H}^1_0(D) \to L^2_*(D)$ is continuous and surjective. (*Hint*: construct $\boldsymbol{v} \in \boldsymbol{H}^1_0(D)$ such that $\boldsymbol{v} = \nabla q + \nabla\times\psi$, where q solves a Poisson problem, ψ solves a biharmonic problem, and $\nabla\times\psi := (\partial_2\psi, -\partial_1\psi)^\mathsf{T}$.)

Exercise 53.2 (de Rham). Let D be a bounded open set in $\mathbb{R}^d$ and assume that D is star-shaped with respect to an open ball $B \subset D$. Prove that the continuous linear forms on $\boldsymbol{W}^{1,p}_0(D)$ that are zero on $\ker(\nabla\cdot)$ are gradients of functions in $L^{p'}_*(D)$. (*Hint*: use Remark 53.10 and the closed range theorem.)

Exercise 53.3 ($\boldsymbol{L}^2$-estimate). Prove Theorem 53.19 directly, i.e., without invoking Lemma 50.11.

Exercise 53.4 (Projection). Let $(\boldsymbol{V}_{h0}, Q_h)_{h\in\mathcal{H}}$ be a sequence of pairs of finite element spaces. Let $p \in [1,\infty]$ and let $p' \in [1,\infty]$ be s.t. $\frac{1}{p}+\frac{1}{p'}=1$. Let $\Pi^Z_h : Q_h \to Z_h$ be an operator, where Z_h is a finite-dimensional subspace of $L^p(D)$. Assume that there are $\beta_1,\beta_2>0$ such that for all $h\in\mathcal{H}$, $\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|\int_D q_h\nabla\cdot\boldsymbol{v}_h\,\mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \ge \beta_1\|q_h - \Pi^Z_h(q_h)\|_{L^{p'}(D)}$ for all $q_h\in Q_h$ and $\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|\int_D q_h\nabla\cdot\boldsymbol{v}_h\,\mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \ge \beta_2\|q_h\|_{L^{p'}(D)}$ for all $q_h\in Z_h$. (i) Show that Π^Z_h is bounded uniformly w.r.t. $h\in\mathcal{H}$. (ii) Show that the $(\boldsymbol{V}_{h0}, Q_h)$ pair satisfies an inf-sup condition uniformly w.r.t. $h\in\mathcal{H}$.

Exercise 53.5 (Spurious mode for the $(\mathbb{Q}_1,\mathbb{Q}_1)$ pair). (i) Let $\widehat{K} := [0,1]^2$ be the unit square. Let $\widehat{\boldsymbol{a}}_{ij} := (\frac{i}{2},\frac{j}{2})$, for all $i,j\in\{0{:}2\}$. Show that the quadrature $\int_{\widehat{K}} f(\widehat{\boldsymbol{x}})\,\mathrm{d}\widehat{x} \approx \sum_{i,j} w_{ij} f(\widehat{\boldsymbol{a}}_{ij})$, where $w_{ij} := \frac{1}{36}(3i(2-i)+1)(3j(j-2)+1)$ ($w_{ij}:=\frac{1}{36}$ for the four vertices of $\widehat{K}$, $w_{ij}:=\frac{1}{9}$ for the four edge midpoints, and $w_{ij}:=\frac{4}{9}$ at the barycenter of $\widehat{K}$) is exact for all $f\in\mathbb{Q}_2$. (*Hint*: write the $\mathbb{Q}_2$ Lagrange shape functions in tensor-product form and use Simpson's rule in each direction.) (ii) Consider $D := (0,1)^2$ and a mesh composed of $I\times I$ squares, $I\ge 2$. Consider the points $\boldsymbol{a}_{lm} := (\frac{l}{2I},\frac{m}{2I})$ for all $l,m\in\{0{:}2I\}$. Let p_h be the continuous, piecewise bilinear function such that $p_h(\boldsymbol{a}_{2k,2n}) := (-1)^{k+n}$ for all $k,n\in\{0{:}I\}$. Show that p_h is a spurious pressure mode for the $(\mathbb{Q}_1,\mathbb{Q}_1)$ pair (continuous velocity and pressure).

Solution to exercises

Exercise 53.1 ($\nabla\cdot$ is surjective). Let $g \in L^2_*(D)$ and let $q_g \in H^1_*(D) := \{v \in H^1(D) \mid \int_D v\,\mathrm{d}x = 0\}$ be such that

$$\Delta q_g = g, \quad \partial_n q_{g|\partial D} = 0.$$

The elliptic regularity theory for the Laplace operator implies that there is a constant c so that $\|q_g\|_{H^2(D)} \le c\|g\|_{L^2(D)}$. Let $\boldsymbol{n} := (n_1, n_2)^\mathsf{T}$ be the outward unit normal at the boundary and define the unit tangent vector $\boldsymbol{\tau} := (-n_2, n_1)^\mathsf{T}$. Owing to the boundedness statement from Theorem 3.10(iii) applied componentwise to ∇q_g (with $p := 2$), we infer that $\nabla q_{g|\partial D} \in \boldsymbol{H}^{\frac12}(\partial D)$, i.e., $\boldsymbol{\tau}\cdot\nabla q_{g|\partial D} \in H^{\frac12}(\partial D)$ (recall that the boundary of D is of class C^2). Let $\psi \in H^2(D)$ be such that

$$\Delta^2\psi_g = 0, \quad \psi_{g|\partial D} = 0, \quad \partial_n\psi_{g|\partial D} = \boldsymbol{\tau}\cdot\nabla q_{g|\partial D}.$$

Let us show that there is c so that $\|\psi_g\|_{H^2(D)} \le c\|g\|_{L^2(D)}$ for all $g \in L^2_*(D)$. Invoking the surjectivity statement from Theorem 3.16(i), there exists $\phi_g \in H^2(D)$ so that $\phi_{g|\partial D} = 0$ and $\partial_n\phi_{g|\partial D} = \boldsymbol{\tau}\cdot\nabla q_{g|\partial D}$, and there is c s.t. $\|\phi_g\|_{H^2(D)} \le c\|\boldsymbol{\tau}\cdot\nabla q_g\|_{H^{\frac12}(\partial D)}$ for all $g \in L^2_*(D)$. Hence, we have (the value of c can change at each occurrence)

$$\begin{aligned}\|\phi_g\|_{H^2(D)} &\le c\,\|\boldsymbol{\tau}\cdot\nabla q_g\|_{H^{\frac12}(\partial D)} \le c\,\|\nabla q_g\|_{\boldsymbol{H}^{\frac12}(\partial D)}\\ &\le c\,\|q_g\|_{H^2(D)} \le c\,\|g\|_{L^2(D)}.\end{aligned}$$

This shows that $\|\phi_g\|_{H^2(D)} \le c\|g\|_{L^2(D)}$. The definitions of ψ_g and ϕ_g imply that

$$\Delta^2(\psi_g - \phi_g) = -\Delta^2\phi_g, \quad (\psi_g - \phi_g)_{|\partial D} = 0, \quad \partial_n(\psi_g - \phi_g)_{|\partial D} = 0.$$

The solution to this problem is such that $\|\Delta(\psi_g - \phi)\|_{L^2(D)} \le \|\Delta\phi_g\|_{L^2(D)}$. This indeed results from

$$\begin{aligned}\|\Delta(\psi_g - \phi_g)\|^2_{L^2(D)} &= (\Delta(\psi_g - \phi_g), \Delta(\psi_g - \phi_g))_{L^2(D)}\\ &= (\Delta^2(\psi_g - \phi_g), \psi_g - \phi_g)_{L^2(D)}\\ &= -(\Delta^2\phi_g, \psi_g - \phi_g)_{L^2(D)} = -(\Delta\phi_g, \Delta(\psi_g - \phi_g))_{L^2(D)},\end{aligned}$$

and the assertion follows from the Cauchy–Schwarz inequality. Moreover, since $(\psi_g - \phi_g)_{|\partial D} = 0$, the elliptic regularity theory implies that

$$\|\psi_g - \phi_g\|_{H^2(D)} \le c\,\|\Delta(\psi_g - \phi_g)\|_{L^2(D)} \le c\,\|\Delta\phi_g\|_{L^2(D)} \le c\,\|g\|_{L^2(D)}.$$

Invoking the triangle inequality and the bound $\|\phi_g\|_{H^2(D)} \le c\|g\|_{L^2(D)}$ shows that there is c, uniform w.r.t. $g \in L^2_*(D)$, such that $\|\psi_g\|_{H^2(D)} \le c\|g\|_{L^2(D)}$. Let us now consider the field

$$\boldsymbol{v}_g := \nabla q_g + \nabla\times\psi_g,$$

where $\nabla\times\psi_g = (\partial_2\psi_g, -\partial_1\psi_g)^\mathsf{T}$. We have $\nabla\cdot\boldsymbol{v}_g = g$ and

$$\boldsymbol{v}_g\cdot\boldsymbol{n} = \partial_n q_g + (n_1\partial_2\psi_g - n_2\partial_1\psi_g) = 0 + \boldsymbol{\tau}\cdot\nabla\psi_g = \partial_\tau\psi_g = 0.$$

Moreover, we have

$$\boldsymbol{v}_g\cdot\boldsymbol{\tau} = \boldsymbol{\tau}\cdot\nabla q_g + (-n_2\partial_2\psi_g - n_1\partial_1\psi_g) = \boldsymbol{\tau}\cdot\nabla q_g - \partial_n\psi_g = 0.$$

In conclusion, $\nabla\cdot\boldsymbol{v}_g = g$, $\boldsymbol{v}_{g|\partial D} = \boldsymbol{0}$, and there is c, uniform w.r.t. $g \in L^2_*(D)$, s.t. $\|\boldsymbol{v}_g\|_{\boldsymbol{H}^1(D)} \le \|q_g\|_{H^2(D)} + \|\psi_g\|_{H^2(D)} \le c\|g\|_{L^2(D)}$.

Exercise 53.2 (de Rham). Consider the weak gradient operator $\nabla : L^{p'}_*(D) \to \boldsymbol{W}^{-1,p'}(D)$. One readily sees that $-\nabla = (\nabla\cdot)^*$. Since $\nabla\cdot$ is surjective owing to Remark 53.10, the closed range theorem implies that $[\ker(\nabla\cdot)]^\perp = \mathrm{im}(\nabla)$.

Exercise 53.3 ($\boldsymbol{L}^2$-estimate). Let $(\boldsymbol{\xi},\phi) \in \boldsymbol{V}_{\mathrm{d}}\times Q$ be the solution to the adjoint problem with source term $\boldsymbol{u}-\boldsymbol{u}_h$, i.e.,

$$a(\boldsymbol{v},\boldsymbol{\xi}) + b(\boldsymbol{v},\phi) = \int_D (\boldsymbol{u}-\boldsymbol{u}_h)\cdot\boldsymbol{v}\,\mathrm{d}x, \quad b(\boldsymbol{\xi},q) = 0, \qquad \forall (\boldsymbol{v},q) \in \boldsymbol{V}_{\mathrm{d}}\times Q.$$

Taking $\boldsymbol{v} := \boldsymbol{u}-\boldsymbol{u}_h$, $q := p-p_h$, and using the Galerkin orthogonality property, i.e., $a(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{\xi}_h)+b(\boldsymbol{\xi}_h,p-p_h)=0$ for all $\boldsymbol{\xi}_h \in \boldsymbol{V}_{h0}$ and $b(\boldsymbol{u}-\boldsymbol{u}_h,\phi_h)=0$ for all $\phi_h \in Q_h$, we obtain

$$\begin{aligned}
\|\boldsymbol{u}-\boldsymbol{u}_h\|^2_{\boldsymbol{L}^2(D)} &= a(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{\xi}) + b(\boldsymbol{u}-\boldsymbol{u}_h,\phi)\\
&= a(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{\xi}-\boldsymbol{\xi}_h) - b(\boldsymbol{\xi}_h,p-p_h) + b(\boldsymbol{u}-\boldsymbol{u}_h,\phi-\phi_h)\\
&= a(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{\xi}-\boldsymbol{\xi}_h) + b(\boldsymbol{\xi}-\boldsymbol{\xi}_h,p-p_h) + b(\boldsymbol{u}-\boldsymbol{u}_h,\phi-\phi_h)\\
&\le \|a\|\,|\boldsymbol{u}-\boldsymbol{u}_h|_{\boldsymbol{H}^1(D)}|\boldsymbol{\xi}-\boldsymbol{\xi}_h|_{\boldsymbol{H}^1(D)}\\
&\quad + \|b\|\,|\boldsymbol{\xi}-\boldsymbol{\xi}_h|_{\boldsymbol{H}^1(D)}\|p-p_h\|_{L^2(D)} + \|b\|\,|\boldsymbol{u}-\boldsymbol{u}_h|_{\boldsymbol{H}^1(D)}\|\phi-\phi_h\|_{L^2(D)}.
\end{aligned}$$

We infer that

$$\begin{aligned}
\|\boldsymbol{u}-\boldsymbol{u}_h\|^2_{\boldsymbol{L}^2(D)} \le \Big(|\boldsymbol{u}-\boldsymbol{u}_h|_{\boldsymbol{H}^1(D)} + \frac{\|b\|}{\|a\|}\|p-p_h\|_{L^2(D)}\Big)\\
\times \Big(\inf_{\boldsymbol{\xi}_h\in\boldsymbol{V}_{h0}} \|a\|\,|\boldsymbol{\xi}-\boldsymbol{\xi}_h|_{\boldsymbol{H}^1(D)} + \inf_{\phi_h\in Q_h}\|b\|\,\|\phi-\phi_h\|_{L^2(D)}\Big),
\end{aligned}$$

and the approximation properties of finite elements imply that

$$\begin{aligned}
\|\boldsymbol{u}-\boldsymbol{u}_h\|^2_{\boldsymbol{L}^2(D)} \le c\Big(\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{H}^1(D)} + \frac{\|b\|}{\|a\|}\|p-p_h\|_{L^2(D)}\Big)\\
\times h^s(\|a\|\,|\boldsymbol{\xi}|_{\boldsymbol{H}^{1+s}(D)} + \|b\|\,|\phi|_{H^s(D)}).
\end{aligned}$$

The conclusion follows from the regularity pickup estimate

$$\|a\|\,|\boldsymbol{\xi}_h|_{\boldsymbol{H}^{1+s}(D)} + \|b\|\,|\phi|_{H^s(D)} \le c\,\ell_D^{1-s}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)}.$$

Exercise 53.4 (Projection). Let q_h be an arbitrary function in Q_h.
(i) Hölder's inequality implies that

$$c\,\|q_h\|_{L^{p'}(D)} \ge \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|\int_D q_h\nabla\cdot\boldsymbol{v}_h\,\mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \ge \beta_1\|q_h - \Pi_h^Z(q_h)\|_{L^{p'}(D)}.$$

Hence, $\|\Pi_h^Z(q_h)\|_{L^{p'}(D)} \le \|\Pi_h^Z(q_h)-q_h\|_{L^{p'}(D)} + \|q_h\|_{L^{p'}(D)} \le (\frac{c}{\beta_1}+1)\|q_h\|_{L^{p'}(D)}$.
(ii) We have

$$\begin{aligned}
\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|\int_D q_h\nabla\cdot\boldsymbol{v}_h\,\mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} &\ge \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|\int_D \Pi_h^Z(q_h)\nabla\cdot\boldsymbol{v}_h\,\mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} - \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|\int_D (q_h-\Pi_h^Z(q_h))\nabla\cdot\boldsymbol{v}_h\,\mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}}\\
&\ge \beta_2\|\Pi_h^Z(q_h)\|_{L^{p'}(D)} - c\,\|q_h-\Pi_h^Z(q_h)\|_{L^{p'}(D)},
\end{aligned}$$

where we used Hölder's inequality and $\|\nabla\cdot\boldsymbol{v}_h\|_{L^p(D)} \le c|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}$ to bound the last term. But we also have

$$\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|\int_D q_h\nabla\cdot\boldsymbol{v}_h\,\mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \ge \beta_1\|q_h-\Pi_h^Z(q_h)\|_{L^{p'}(D)}.$$

The above two inequalities imply that

$$\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}}\frac{|\int_D q_h\nabla\cdot\boldsymbol{v}_h\,\mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}}\geq\frac{\beta_1\beta_2}{c+\beta_1}\|\Pi_h^Z(q_h)\|_{L^{p'}(D)}.$$

This, in turn, implies that

$$\left(\frac{c+\beta_1}{\beta_1\beta_2}+\frac{1}{\beta_1}\right)\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}}\frac{\int_D q_h\nabla\cdot\boldsymbol{v}_h\,\mathrm{d}x}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}}\geq\|\Pi_h^Z(q_h)\|_{L^{p'}(D)}+\|q_h-\Pi_h^Z(q_h)\|_{L^{p'}(D)}\geq\|q_h\|_{L^{p'}(D)},$$

which shows that $(\boldsymbol{V}_{h0},Q_h)$ satisfies a uniform inf-sup condition.

Exercise 53.5 (Spurious mode for the $(\mathbb{Q}_1,\mathbb{Q}_1)$ pair). (i) Let $\widehat{\theta}_{ij}$ be the $\mathbb{Q}_2$ Lagrange shape function associated with the node $\widehat{\boldsymbol{a}}_{ij}:=(\frac{i}{2},\frac{j}{2})$, $i,j\in\{0{:}2\}$. This shape function can be represented as $\widehat{\theta}_{ij}(\widehat{\boldsymbol{x}}):=p_i(\widehat{x}_1)q_j(\widehat{x}_2)$, where $\widehat{\boldsymbol{x}}:=(\widehat{x}_1,\widehat{x}_2)^\mathsf{T}$ and p_i, q_j are univariate quadratic polynomials. Using Simpson's rule yields

$$\begin{aligned}\int_{\widehat{K}}\widehat{\theta}_{ij}(\widehat{\boldsymbol{x}})\,\mathrm{d}\widehat{x}&=\left(\int_0^1 p_i(\widehat{x}_1)\,\mathrm{d}\widehat{x}_1\right)\left(\int_0^1 q_j(\widehat{x}_2)\,\mathrm{d}\widehat{x}_2\right)\\&=\frac{1}{36}(p_i(0)+4p_i(\tfrac12)+p_i(1))(q_j(0)+4q_j(\tfrac12)+q_j(1))\\&=\frac{1}{36}\left(\sum_{l\in\{0:2\}}(3l(2-l)+1)p_i(\tfrac{l}{2})\right)\left(\sum_{m\in\{0:2\}}(3m(2-m)+1)q_j(\tfrac{m}{2})\right)\\&=\sum_{l,m\in\{0:2\}}w_{lm}\widehat{\theta}_{ij}(\widehat{\boldsymbol{a}}_{lm})=w_{ij},\end{aligned}$$

where $w_{lm}:=\frac{1}{36}(3l(2-l)+1)(3m(2-m)+1)$. The conclusion follows readily since $(\widehat{\theta}_{ij})_{i,j\in\{0:2\}}$ is a basis of $\mathbb{Q}_2$.
(ii) Let us first observe that $p_h(\boldsymbol{a}_{2k+1,2n})=\frac12((-1)^{k+n}+(-1)^{k+1+n})=0$ for all $k\in\{0{:}I{-}1\}$ and all $n\in\{0{:}I\}$. Similarly, $p_h(\boldsymbol{a}_{2k,2n+1})=0$ for all $k\in\{0{:}I\}$ and all $n\in\{0{:}I{-}1\}$, and $p_h(\boldsymbol{a}_{2k+1,2n+1})=0$ for all $k\in\{0{:}I-1\}$ and all $n\in\{0{:}I{-}1\}$. Let $\boldsymbol{\varphi}_{ij}$ be a global shape function (for the $\mathbb{Q}_1$ Lagrange element) associated with the node $\boldsymbol{a}_{2i,2j}$, for all $i,j\in\{1{:}I{-}1\}$ (recall that $I\geq2$ by assumption). It suffices to show that $\int_D(\nabla\cdot\boldsymbol{\varphi}_{ij})p_h\,\mathrm{d}x=0$. Since $(\nabla\cdot\boldsymbol{\varphi}_{ij})p_h$ is piecewise in the polynomial space $\mathbb{Q}_{2,2}$, the function $\boldsymbol{\varphi}_{ij}$ is supported in the four cells sharing $\boldsymbol{a}_{2i,2j}$ and since $\nabla\cdot\boldsymbol{\varphi}_{ij}(\boldsymbol{a}_{2(i\pm1),2(j\pm1)})=0$, the quadrature from Step (i) yields

$$\begin{aligned}\int_D(\nabla\cdot\boldsymbol{\varphi}_{ij})p_h\,\mathrm{d}x=\frac{(-1)^{i+j}}{36N^2}\Big(&4(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2i,2j})-(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2(i-1),2j})\\&-(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2(i+1),2j})-(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2i,2(j-1)})-(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2i,2(j+1)})\Big).\end{aligned}$$

Symmetry arguments show that $(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2i,2j})=0$, $(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2(i-1),2j})+(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2(i+1),2j})=0$ and $(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2i,2(j-1)})+(\nabla\cdot\boldsymbol{\varphi}_{ij})(\boldsymbol{a}_{2i,2(j+1)})=0$. Hence, $\int_D(\nabla\cdot\boldsymbol{\varphi}_{ij})p_h\,\mathrm{d}x=0$ for all $i,j\in\{1{:}I{-}1\}$. This shows that p_h is a spurious pressure mode.

Chapter 54

Stokes equations: Stable pairs (I)

Exercises

Exercise 54.1 (Mini element). Show that the Fortin operator $\boldsymbol{\Pi}_h$ constructed in the proof of Lemma 54.5 is of the form $\boldsymbol{\Pi}_h(\boldsymbol{v}) \coloneqq \boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}) + \sum_{K\in\mathcal{T}_h}\sum_{i\in\{1:d\}}\gamma_K^i(\boldsymbol{v})b_K\boldsymbol{e}_i$, for some coefficients $\gamma_K^i(\boldsymbol{v})$ to be determined. Here, $\{\boldsymbol{e}_i\}_{i\in\{1:d\}}$ is the canonical Cartesian basis of $\mathbb{R}^d$.

Exercise 54.2 (Bubble⇔Stabilization). Consider the mini element defined in §54.2 and assume that the viscosity μ is constant over D. Recall that $\boldsymbol{V}_{h0} \coloneqq \boldsymbol{V}_{h0}^1 \oplus \boldsymbol{B}_h$ and $Q_h \coloneqq P_1^{\mathrm{g}}(\mathcal{T}_h)\cap L^2_*(D)$ with $\boldsymbol{V}_{h0}^1 \coloneqq \boldsymbol{\mathbb{P}}_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$. Let $(\boldsymbol{u}_h,p_h)$ be the solution to the discrete Stokes problem (53.14). (i) Show that $a(\boldsymbol{v}_h,\boldsymbol{b}_h)=0$ for all $\boldsymbol{v}_h\in\boldsymbol{V}_{h0}^1$ and all $\boldsymbol{b}_h\in\boldsymbol{B}_h$. (ii) Set $\boldsymbol{u}_h \coloneqq \boldsymbol{u}_h^1+\boldsymbol{u}_h^b\in\boldsymbol{V}_{h0}$. Show that

$$a(\boldsymbol{u}_h^1,\boldsymbol{v}_h)+b(\boldsymbol{v}_h,p_h)=F(\boldsymbol{v}_h),\qquad \forall\boldsymbol{v}_h\in\boldsymbol{V}_{h0}^1. \tag{54.1}$$

(iii) Let $b_K \coloneqq \widehat{b}\circ\boldsymbol{T}_K$ be the bubble function on $K\in\mathcal{T}_h$. Let $\{\boldsymbol{e}_i\}_{i\in\{1:d\}}$ be the canonical Cartesian basis of $\mathbb{R}^d$. Let $\mathcal{S}^K\in\mathbb{R}^{d\times d}$ be defined by $\mathcal{S}_{ij}^K \coloneqq \frac{1}{\int_K b_K\,\mathrm{d}x}a(b_K\boldsymbol{e}_j,b_K\boldsymbol{e}_i)$ for all $i,j\in\{1:d\}$. Let $\boldsymbol{u}_{h|K}^b \coloneqq \sum_{i\in\{1:d\}}c_K^i\boldsymbol{e}_ib_K$. Show that $\boldsymbol{c}_K=(\mathcal{S}^K)^{-1}(\boldsymbol{F}_K-\nabla p_{h|K})$, where $F_K^i \coloneqq \frac{1}{\int_K b_K\,\mathrm{d}x}F(b_K\boldsymbol{e}_i)$, for all $i\in\{1:d\}$. (iv) Set $c_h(p_h,q_h) \coloneqq \sum_{K\in\mathcal{T}_h}\nabla q_{h|K}(\mathcal{S}^K)^{-1}\nabla p_{h|K}\int_K b_K\,\mathrm{d}x$ and $R_h(q_h) \coloneqq \sum_{K\in\mathcal{T}_h}\nabla q_{h|K}(\mathcal{S}^K)^{-1}\boldsymbol{F}_K\int_K b_K\,\mathrm{d}x$. Show that the mass conservation equation becomes

$$b(\boldsymbol{u}_h^1,q_h)-c_h(p_h,q_h)=G(q_h)-R_h(q_h),\qquad\forall q_h\in Q_h. \tag{54.2}$$

Note: since $(\mathcal{S}_K)^{-1}$ scales like $\mu^{-1}h_K^2$, $c_h(p_h,q_h)$ behaves like $\sum_{K\in\mathcal{T}_h}\frac{h_K^2}{\mu}\int_K\nabla q_h\cdot\nabla p_h\,\mathrm{d}x$, and $R_h(q_h)$ scales like $\sum_{K\in\mathcal{T}_h}\frac{h_K^2}{\mu}\int_K\nabla q_{h|K}\cdot F_K\,\mathrm{d}x$. This shows that, once the bubbles are eliminated, the system (54.1)-(54.2) is equivalent to a stabilized form of the Stokes system for the $(\mathbb{P}_1,\mathbb{P}_1)$ pair; see Chapters 62 and 63.

Exercise 54.3 (Singular vertex). Let $K\subset\mathbb{R}^2$ be a quadrangle and let $\boldsymbol{z}$ be the intersection of the two diagonals of K. Let $K_1,\dots,K_4$ be the four triangles formed by dividing K along its two diagonals (assume that $K_1\cap K_3=\{\boldsymbol{z}\}$ and $K_2\cap K_4=\{\boldsymbol{z}\}$). (i) Let ϕ be a scalar field continuous over K and of class C^1 over the triangles $K_1,\dots,K_4$. Prove that $\sum_{i\in\{1:4\}}(-1)^i\boldsymbol{n}\cdot\nabla\phi_{|K_i}(\boldsymbol{z})=0$ for every unit vector $\boldsymbol{n}$. (ii) Let $\boldsymbol{v}$ be a vector field continuous over K and of class C^1 over the triangles $K_1,\dots,K_4$. Prove that $\sum_{i\in\{1:4\}}(-1)^i\nabla\cdot\boldsymbol{v}_{|K_i}(\boldsymbol{z})=0$. (iii) Assume that $\boldsymbol{v}$ is linear over each triangle. Show that the four equations $\int_{K_i}\nabla\cdot\boldsymbol{v}\,\mathrm{d}x=0$ for all $i\in\{1:4\}$ are linearly dependent.

Exercise 54.4 ($\mathbb{P}_1$-iso-$\mathbb{P}_2,\mathbb{P}_1$). Consider the setting of Lemma 54.12 with the ($\mathbb{P}_1$-iso-$\mathbb{P}_2,\mathbb{P}_1$) pair in dimension three. (i) Let $K\in\mathcal{T}_h$. Let $\mathcal{V}_K$ be the set of the vertices of K. Let $\mathcal{M}_K$ be the midpoints of the six edges of K. Let $\mathcal{M}_K^1$ be the set of the two midpoints that are connected to create the 8 new tetrahedra. Let $\mathcal{M}_K^2$ be the set of the remaining midpoints. Let $\boldsymbol{V}_{h0}$ be the $\mathbb{P}_1$ velocity space based of $\mathcal{T}_{h/2}$. Find the coefficients α,β,γ so that the following quadrature is exact for all $\boldsymbol{w}_h\in V_{h0}$: $\int_K \boldsymbol{w}_h\,\mathrm{d}x = |K|(\alpha\sum_{\boldsymbol{z}\in\mathcal{V}_K}\boldsymbol{w}_h(\boldsymbol{z})+\beta\sum_{\boldsymbol{m}\in\mathcal{M}_K^1}\boldsymbol{w}_h(\boldsymbol{m})+\gamma\sum_{\boldsymbol{m}\in\mathcal{M}_K^2}\boldsymbol{w}_h(\boldsymbol{m}))$. (*Hint*: on a tetrahedron K' with vertices $\{\boldsymbol{z}'\}_{\boldsymbol{z}'\in\mathcal{V}_{K'}}$, the quadrature $\int_{K'}\boldsymbol{w}_h\,\mathrm{d}x=|K'|\sum_{\boldsymbol{z}'\in\mathcal{V}_{K'}}\frac14\boldsymbol{w}_h(\boldsymbol{z}')$ is exact on $\mathbb{P}_1$.) (ii) Prove Lemma 54.12 for the ($\mathbb{P}_1$-iso-$\mathbb{P}_2,\mathbb{P}_1$) pair in dimension three for all $p\in(1,\infty)$. (*Hint*: adapt the proof of Lemma 54.8.)

Solution to exercises

Exercise 54.1 (Mini element). A direct calculation shows that

$$\gamma_K^i=\frac{\int_K(v_i-\mathcal{I}_{h0}^{\mathrm{g,av}}(v_i))\,\mathrm{d}x}{\int_K b_K\,\mathrm{d}x},$$

for all $i\in\{1\!:\!d\}$ and all $K\in\mathcal{T}_h$.

Exercise 54.2 (Bubble⇔Stabilization). (i) Since the mesh is affine, the function $\mathrm{e}(\boldsymbol{v}_h)$ is linear over each cell $K\in\mathcal{T}_h$ for all $\boldsymbol{v}_h\in\boldsymbol{V}_{h0}^1$. Hence, we have

$$\begin{aligned}\frac{1}{2\mu}a(\boldsymbol{v}_h,\boldsymbol{b}_h)&=\sum_K\int_K\mathrm{e}(\boldsymbol{v}_h)\!:\!\mathrm{e}(\boldsymbol{b}_h)\,\mathrm{d}x=\sum_K\int_K\mathrm{e}(\boldsymbol{v}_h)\!:\!\nabla\boldsymbol{b}_h\,\mathrm{d}x\\&=\sum_K-\int_K\nabla\cdot(\mathrm{e}(\boldsymbol{v}_h))\cdot\boldsymbol{b}_h\,\mathrm{d}x=0.\end{aligned}$$

(ii) Since the bilinear form a is symmetric, the above argument gives $a(\boldsymbol{u}_h^b,\boldsymbol{v}_h)=a(\boldsymbol{v}_h,\boldsymbol{u}_h^b)=0$. The assertion follows readily.
(iii) Let $K\in\mathcal{T}$. Testing the momentum conservation equation against the function $b_K\boldsymbol{e}_i$, we obtain

$$\begin{aligned}\sum_{j\in\{1:d\}}a(b_K\boldsymbol{e}_j,b_K\boldsymbol{e}_i)c_K^j&=a(\boldsymbol{u}_h^b,b_K\boldsymbol{e}_i)=a(\boldsymbol{u}_h,b_K\boldsymbol{e}_i)\\&=-b(b_K\boldsymbol{e}_i,p_h)+F(b_K\boldsymbol{e}_i)\\&=-\int_K b_K\partial_i p_h\,\mathrm{d}x+F(b_K\boldsymbol{e}_i).\end{aligned}$$

Dividing by $\int_K b_K\,\mathrm{d}x$, we obtain $\mathcal{S}^K\boldsymbol{c}_K=(-\nabla p_{h|K}+\boldsymbol{F}_K)$. This proves the assertion.
(iv) The mass conservation equation gives for all $q_h\in Q_h$,

$$\begin{aligned}G(q_h)-b(\boldsymbol{u}_h^1,q_h)=b(\boldsymbol{u}_h^b,q_h)&=\sum_{K\in\mathcal{T}_h}\int_K b_K\boldsymbol{c}_K\cdot\nabla q_h\,\mathrm{d}x\\&=\sum_{K\in\mathcal{T}_h}\nabla q_{h|K}(\mathcal{S}^K)^{-1}(-\nabla p_{h|K}+\boldsymbol{F}_K)\int_K b_K\,\mathrm{d}x.\end{aligned}$$

Using the notation

$$c_h(p_h,q_h) := \sum_{K\in\mathcal{T}_h} \nabla q_{h|K}(\mathcal{S}^K)^{-1}\nabla p_{h|K}\int_K b_K\,\mathrm{d}x,$$
$$R_h(q_h) := \sum_{K\in\mathcal{T}_h} \nabla q_{h|K}(\mathcal{S}^K)^{-1}\boldsymbol{F}_K\int_K b_K\,\mathrm{d}x,$$

the mass conservation equation becomes

$$b(\boldsymbol{u}_h^1,q_h) - c_h(p_h,q_h) = G(q_h) - R_h(q_h).$$

Exercise 54.3 (Singular vertex). Without loss of generality, assume that the enumeration of the triangles $K_1,\dots,K_4$ is done counter-clockwise. Let $\boldsymbol{\tau}_1$ be a unit vector aligned with the diagonal, say Δ_1, separating K_1 and K_4, and K_2 and K_3. Likewise, let $\boldsymbol{\tau}_2$ be a unit vector aligned with the diagonal, say Δ_2, separating K_1 and K_2, and K_3 and K_4. Let $\phi\in C^0(K;\mathbb{R})$ and assume that $\phi_i := \phi_{|K_i}\in C^1(K_i;\mathbb{R})$. We have

$$\boldsymbol{\tau}_1\cdot\nabla\phi_1 = \boldsymbol{\tau}_1\cdot\nabla\phi_4,\quad \boldsymbol{\tau}_1\cdot\nabla\phi_3 = \boldsymbol{\tau}_1\cdot\nabla\phi_2,\quad \text{on } \Delta_1,$$
$$\boldsymbol{\tau}_2\cdot\nabla\phi_4 = \boldsymbol{\tau}_2\cdot\nabla\phi_3,\quad \boldsymbol{\tau}_2\cdot\nabla\phi_2 = \boldsymbol{\tau}_2\cdot\nabla\phi_1,\quad \text{on } \Delta_2.$$

Since $\boldsymbol{z}\in\Delta_1\cap\Delta_2$, we infer that

$$-\boldsymbol{\tau}_1\cdot\nabla\phi_1(\boldsymbol{z}) + \boldsymbol{\tau}_1\cdot\nabla\phi_2(\boldsymbol{z}) - \boldsymbol{\tau}_1\cdot\nabla\phi_3(\boldsymbol{z}) + \boldsymbol{\tau}_1\cdot\nabla\phi_4(\boldsymbol{z}) = 0,$$
$$-\boldsymbol{\tau}_2\cdot\nabla\phi_1(\boldsymbol{z}) + \boldsymbol{\tau}_2\cdot\nabla\phi_2(\boldsymbol{z}) - \boldsymbol{\tau}_2\cdot\nabla\phi_3(\boldsymbol{z}) + \boldsymbol{\tau}_2\cdot\nabla\phi_4(\boldsymbol{z}) = 0.$$

Let $\boldsymbol{n} = \alpha\boldsymbol{\tau}_1 + \beta\boldsymbol{\tau}_2$ be any unit vector in $\mathbb{R}^2$ (recall that $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ are linearly independent). Combining the above two equations, we infer that

$$-\boldsymbol{n}\cdot\nabla\phi_1(\boldsymbol{z}) + \boldsymbol{n}\cdot\nabla\phi_2(\boldsymbol{z}) - \boldsymbol{n}\cdot\nabla\phi_3(\boldsymbol{z}) + \boldsymbol{n}\cdot\nabla\phi_4(\boldsymbol{z}) = 0,$$

i.e., $\sum_{i\in\{1:4\}}(-1)^i\boldsymbol{n}\cdot\nabla\phi_i(\boldsymbol{z}) = 0$.
(ii) We can now apply this result to the Cartesian components of the vector field $\boldsymbol{v}$ using $\boldsymbol{n} = \boldsymbol{e}_x$ and $\boldsymbol{n} = \boldsymbol{e}_y$. Let us set $\boldsymbol{v}_i := \boldsymbol{v}_{|K_i}$ and let v_i^x, v_i^y be the two Cartesian components of $\boldsymbol{v}_i$. We have

$$0 = \sum_{i\in\{1:4\}}(-1)^i\partial_x v_i^x(\boldsymbol{z}) + \sum_{i\in\{1:4\}}(-1)^i\partial_y v_i^y(\boldsymbol{z}) = \sum_{i\in\{1:4\}}(-1)^i\nabla\cdot\boldsymbol{v}_i(\boldsymbol{z}).$$

(iii) If we assume that $\boldsymbol{v}$ is piecewise linear, then $\nabla\cdot\boldsymbol{v}_i$ is constant over K_i, i.e., $\int_{K_i}\nabla\cdot\boldsymbol{v}_i\,\mathrm{d}x = |K_i|\nabla\cdot\boldsymbol{v}_i(\boldsymbol{z})$. Hence, $\sum_{i\in\{1:4\}}(-1)^i|K_i|^{-1}\int_{K_i}\nabla\cdot\boldsymbol{v}_i\,\mathrm{d}x = 0$, which shows that the four equations $\int_{K_i}\nabla\cdot\boldsymbol{v}_i\,\mathrm{d}x = 0$ for all $i\in\{1{:}4\}$ are linearly dependent.

Exercise 54.4 ($\mathbb{P}_1$-iso-$\mathbb{P}_2,\mathbb{P}_1$). We assume that $d = 3$.
(i) Let $\mathcal{T}_{K/8}$ be the collection of the eight tetrahedra created by dividing K. Notice that the eight tetrahedra are not all similar, but they all have the same volume $\frac{1}{8}|K|$. The following holds true for all $\boldsymbol{w}_h\in\boldsymbol{V}_{h0}$:

$$\int_K\boldsymbol{w}_h\,\mathrm{d}x = \frac{|K|}{8}\frac{1}{4}\sum_{K'\in\mathcal{T}_{K/8}}\Big(\sum_{\boldsymbol{z}\in\mathcal{V}_{K'}}\boldsymbol{w}_h(\boldsymbol{z})\Big).$$

We observe that the vertices in $\mathcal{V}_K$ belong to only one cell in $\mathcal{T}_{K/8}$, the midpoints in $\mathcal{M}_K^1$ belong to $2+4=6$ cells in $\mathcal{T}_{K/8}$, and the midpoints in $\mathcal{M}_K^2$ belong to $2+2=4$ cells in $\mathcal{T}_{K/8}$. Rearranging the summations leads to the quadrature

$$\int_K \boldsymbol{w}_h \,\mathrm{d}x = \frac{|K|}{32}\Big(\sum_{\boldsymbol{z}\in\mathcal{V}_K} \boldsymbol{w}_h(\boldsymbol{z}) + 6 \sum_{\boldsymbol{m}\in\mathcal{M}_K^1} \boldsymbol{w}_h(\boldsymbol{m}) + 4 \sum_{\boldsymbol{m}\in\mathcal{M}_K^2} \boldsymbol{w}_h(\boldsymbol{m})\Big).$$

Hence, we have $\alpha = \frac{1}{32}$, $\beta = \frac{3}{16}$, $\gamma = \frac{1}{8}$.
(ii) Let $p \in (1,\infty)$. We proceed as in the proof of Lemma 54.8. Let $\mathcal{T}_h$ be the pressure mesh and $\mathcal{T}_{h/2}$ be the velocity mesh. Let us number all the internal mesh edges of $\mathcal{T}_h$ from 1 to $N_{\mathrm{e}}^{\mathrm{i}}$. Consider an oriented edge E_i with $i \in \{1{:}N_{\mathrm{e}}^{\mathrm{i}}\}$, and denote its two endpoints by $\boldsymbol{z}_i^{\pm}$ and its midpoint by $\boldsymbol{m}_i$. Set $l_i := \|\boldsymbol{z}_i^+ - \boldsymbol{z}_i^-\|_{\ell^2}$ and $\boldsymbol{\tau}_i := l_i^{-1}(\boldsymbol{z}_i^+ - \boldsymbol{z}_i^-)$, so that l_i is the length of E_i and $\boldsymbol{\tau}_i$ is the unit tangent vector orienting E_i. Let q_h be a function in Q_h and let sgn be the sign function. Let $\boldsymbol{v}_h \in \boldsymbol{V}_{h0}$ be (uniquely) defined by prescribing its global degrees of freedom in $\boldsymbol{V}_{h0}$ as follows:

$$\begin{cases} \boldsymbol{v}_h(\boldsymbol{a}_j) := \boldsymbol{0} & \text{if } \boldsymbol{a}_j \text{ is a mesh vertex,} \\ \boldsymbol{v}_h(\boldsymbol{m}_i) := -l_i^{p'} \operatorname{sgn}(\partial_{\tau_i} q_h) |\partial_{\tau_i} q_h|^{p'-1} \boldsymbol{\tau}_i & \text{if } E_i \not\subset \partial D, \\ \boldsymbol{v}_h(\boldsymbol{m}_i) := \boldsymbol{0} & \text{if } E_i \subset \partial D, \end{cases}$$

where $\partial_{\tau_i} q_h := \boldsymbol{\tau}_i \cdot \nabla q_h$ denotes the tangential derivative of q_h along the oriented edge E_i. Note that $\boldsymbol{v}_h(\boldsymbol{m}_i)$ depends only on the values of q_h on E_i. Let $K \in \mathcal{T}_h$. Using that $\nabla q_{h|K}$ is constant over each cell in $\mathcal{T}_h$, and since Q_h is H^1-conforming, we infer that

$$\begin{aligned} \int_D q_h \nabla\cdot\boldsymbol{v}_h \,\mathrm{d}x &= -\int_D \boldsymbol{v}_h \cdot \nabla q_h \,\mathrm{d}x = -\sum_{K\in\mathcal{T}_h} \nabla q_{h|K} \cdot \int_K \boldsymbol{v}_h \,\mathrm{d}x \\ &= -\sum_{K\in\mathcal{T}_h} |K| \Big(\sum_{\boldsymbol{m}_i \in \mathcal{M}_K^1} \tfrac{3}{16} \boldsymbol{v}_h(\boldsymbol{m}_i)\cdot\nabla q_h(\boldsymbol{m}_i) + \sum_{\boldsymbol{m}_i \in \mathcal{M}_K^2} \tfrac{1}{8} \boldsymbol{v}_h(\boldsymbol{m}_i)\cdot\nabla q_h(\boldsymbol{m}_i) \Big) \\ &\ge \sum_{K\in\mathcal{T}_h} \tfrac{1}{8}|K| \sum_{\boldsymbol{m}_i\in K} |\partial_{\tau_i} q_h(\boldsymbol{m}_i)|^{p'} l_i^{p'} \ge c \sum_{K\in\mathcal{T}_h} h_K^{p'} \|\nabla q_h\|_{\boldsymbol{L}^{p'}(K)}^{p'}. \end{aligned}$$

The last inequality results from the fact that $l_i \ge c h_K$ owing to the regularity of the mesh sequence, and that every tetrahedron $K \in \mathcal{T}_h$ has at least three edges in D, i.e., the quantities $|\partial_{\tau_i} q_h(\boldsymbol{m}_i)|$, where $\boldsymbol{m}_i$ spans the midpoints of the edges of K that are not in ∂D, control $\|\nabla q_h\|_{\ell^2}$. Finally, the inverse inequality from Lemma 12.1 (with $r := p$, $l := 1$, $m := 0$) together with Proposition 12.5 implies that for all $K \in \mathcal{T}_h$,

$$|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(K)}^p \le c\, h_K^{-p} |K| \sum_{\boldsymbol{m}\in\mathcal{M}_K} \|\boldsymbol{v}_h(\boldsymbol{m})\|_{\ell^2}^p,$$

and since $l_i \le c h_K$, we have $\|\boldsymbol{v}_h(\boldsymbol{m})\|_{\ell^2} \le c h_K^{p'} \|\nabla q_h\|_{\ell^2}^{p'-1}$. Since $p(p'-1) = p'$, combining these bounds shows that $|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(K)}^p \le c h_K^{p'} \|\nabla q_h\|_{\boldsymbol{L}^{p'}(K)}^{p'}$ for all $K \in \mathcal{T}_h$. This proves that

$$\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|\int_D q_h \nabla\cdot\boldsymbol{v}_h \,\mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \ge c \left(\sum_{K\in\mathcal{T}_h} h_K^{p'} \|\nabla q_h\|_{\boldsymbol{L}^{p'}(K)}^{p'} \right)^{\frac{1}{p'}}.$$

We conclude by applying Lemma 54.3.

Chapter 55

Stokes equations: Stable pairs (II)

Exercises

Exercise 55.1 (Local mass balance). Let $\boldsymbol{u}_h \in \boldsymbol{V}_{h0}$ and $g \in L^2_*(D)$ satisfy $\int_D q_h \nabla{\cdot}\boldsymbol{u}_h \,\mathrm{d}x = \int_D q_h g \,\mathrm{d}x$ for all $q_h \in P^{\mathrm{b}}_{k,*}(\mathcal{T}_h)$. Show that $\int_K (\psi_K^{\mathrm{g}})^{-1}(q) \nabla{\cdot}\boldsymbol{u}_h \,\mathrm{d}x = \int_K (\psi_K^{\mathrm{g}})^{-1}(q) g \,\mathrm{d}x$ for all $q \in \mathbb{P}_{k,d}$ and all $K \in \mathcal{T}_h$ with $\psi_K^{\mathrm{g}}(q) := q \circ \boldsymbol{T}_K$. (*Hint*: use that $\int_D \nabla{\cdot}\boldsymbol{u}_h \,\mathrm{d}x = \int_D g \,\mathrm{d}x = 0$.)

Exercise 55.2 ($(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_0)$). Complete the proof of Lemma 55.8. (*Hint*: to show that the assumption (ii) from Lemma 54.2 is met, prove that $\int_F (\boldsymbol{v} - \boldsymbol{\Pi}_{2h}(\boldsymbol{v})) \,\mathrm{d}s = \boldsymbol{0}$ for all $F \in \mathcal{F}_h^\circ$ using Simpson's quadrature rule; to show that the assumption (iii) is met, show first that $|\boldsymbol{\Pi}_{2h}(\boldsymbol{v})|_{\boldsymbol{W}^{1,p}(K)} \le c h_K^{\frac{1}{p}-1} \sum_{F \in \mathcal{F}_K^\circ} \|\boldsymbol{v}\|_{\boldsymbol{L}^p(F)}$ and then invoke the multiplicative trace inequality (12.16).)

Exercise 55.3 ($(\mathbb{Q}_k, \mathbb{Q}^{\mathrm{b}}_{k-1})$). (i) Justify Lemma 55.23 for $k := 2$ by constructing a counterexample. (*Hint*: given an interior vertex of a uniform Cartesian mesh, consider the patch composed of the four square cells sharing this vertex, and find an oscillating pressure field using (ii) from Exercise 54.3.) (ii) Generalize the argument for all $k \ge 2$.

Exercise 55.4 ($(\mathbb{P}_1^{\mathrm{CR}}, \mathbb{P}^{\mathrm{b}}_0)$). Justify the claim in Remark 55.19. (*Hint*: see the proof of Theorem 36.11.)

Exercise 55.5 ($(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_1)$, HCT mesh). Using the notation from the proof of Lemma 55.14, the goal is to prove that $\mathrm{im}(\widehat{B})^\perp = \mathrm{span}(\mathbb{1}_{\widehat{U}})$. Let $\widehat{\boldsymbol{z}}_1 := (0,0)$, $\widehat{\boldsymbol{z}}_2 := (1,0)$, $\widehat{\boldsymbol{z}}_3 := (0,1)$, $\widehat{\boldsymbol{z}}_4 := (\frac{1}{3}, \frac{1}{3})$. Consider the triangles $\widehat{K}_1 := \mathrm{conv}(\widehat{\boldsymbol{z}}_1, \widehat{\boldsymbol{z}}_2, \widehat{\boldsymbol{z}}_4)$, $\widehat{K}_2 := \mathrm{conv}(\widehat{\boldsymbol{z}}_2, \widehat{\boldsymbol{z}}_3, \widehat{\boldsymbol{z}}_4)$, and $\widehat{K}_3 := \mathrm{conv}(\widehat{\boldsymbol{z}}_3, \widehat{\boldsymbol{z}}_1, \widehat{\boldsymbol{z}}_4)$. Let $p \in P^{\mathrm{b}}_1(\widehat{U})$ with the reference macroelement $\widehat{U} := \{\widehat{K}_1, \widehat{K}_2, \widehat{K}_3\}$, and set

$$
\begin{aligned}
p_1 &:= p_{|\widehat{K}_1}(\widehat{\boldsymbol{z}}_1),\ p_2 := p_{|\widehat{K}_1}(\widehat{\boldsymbol{z}}_2),\ p_3 := p_{|\widehat{K}_1}(\widehat{\boldsymbol{z}}_4),\\
q_1 &:= p_{|\widehat{K}_2}(\widehat{\boldsymbol{z}}_2),\ q_2 := p_{|\widehat{K}_2}(\widehat{\boldsymbol{z}}_3),\ q_3 := p_{|\widehat{K}_2}(\widehat{\boldsymbol{z}}_4),\\
s_1 &:= p_{|\widehat{K}_3}(\widehat{\boldsymbol{z}}_3),\ s_2 := p_{|\widehat{K}_3}(\widehat{\boldsymbol{z}}_1),\ s_3 := p_{|\widehat{K}_3}(\widehat{\boldsymbol{z}}_4).
\end{aligned}
$$

Let $\widehat{\boldsymbol{m}}_{14} := \frac{1}{2}(\widehat{\boldsymbol{z}}_1 + \widehat{\boldsymbol{z}}_4)$, $\widehat{\boldsymbol{m}}_{24} := \frac{1}{2}(\widehat{\boldsymbol{z}}_2 + \widehat{\boldsymbol{z}}_4)$, $\widehat{\boldsymbol{m}}_{34} := \frac{1}{2}(\widehat{\boldsymbol{z}}_3 + \widehat{\boldsymbol{z}}_4)$. Let $\boldsymbol{u} \in \boldsymbol{P}^{\mathrm{g}}_{2,0}(\widehat{U})$ and set $(u_7, v_7)^{\mathsf{T}} := \boldsymbol{u}(\widehat{\boldsymbol{m}}_{14})$, $(u_8, v_8)^{\mathsf{T}} := \boldsymbol{u}(\widehat{\boldsymbol{m}}_{24})$, $(u_9, v_9)^{\mathsf{T}} := \boldsymbol{u}(\widehat{\boldsymbol{m}}_{34})$, $(u_{10}, v_{10})^{\mathsf{T}} := \boldsymbol{u}(\widehat{\boldsymbol{z}}_4)$. (i) Show (or

accept as a fact) that

$$\int_{\widehat{K}_1} p\nabla\cdot\boldsymbol{u}\,\mathrm{d}\widehat{x} = (-u_7+u_8+4v_7+2v_8)p_1 + (-u_7+u_8+v_7+5v_8)p_2 + (-2u_7+2u_8-v_7+v_8+3v_{10})p_3.$$

(*Hint*: compute the $\mathbb{P}_2$ shape functions on $\widehat{K}_1$ associated with the nodes $\widehat{\boldsymbol{m}}_{14}$, $\widehat{\boldsymbol{m}}_{24}$, and $\widehat{\boldsymbol{z}}_4$.) (ii) Let $\boldsymbol{T}_{\widehat{K}_2}:\widehat{K}_1\to\widehat{K}_2$, $\boldsymbol{T}_{\widehat{K}_3}:\widehat{K}_1\to\widehat{K}_3$ be the geometric mappings s.t.

$$\boldsymbol{T}_{\widehat{K}_2}(\widehat{\boldsymbol{x}}) := \widehat{\boldsymbol{z}}_2 + \begin{pmatrix}-1 & -1\\ 1 & 0\end{pmatrix}(\widehat{\boldsymbol{x}}-\widehat{\boldsymbol{z}}_1),\quad \boldsymbol{T}_{\widehat{K}_3}(\widehat{\boldsymbol{x}}) := \widehat{\boldsymbol{z}}_3 + \begin{pmatrix}0 & 1\\ -1 & -1\end{pmatrix}(\widehat{\boldsymbol{x}}-\widehat{\boldsymbol{z}}_1).$$

Verify that $\boldsymbol{T}_{\widehat{K}_i}$ maps the vertices of $\widehat{K}_1$ to the vertices of $\widehat{K}_i$ for $i\in\{2,3\}$. (iii) Compute the contravariant Piola tranformations $\boldsymbol{\psi}^{\mathrm{d}}_{\widehat{K}_2}(\boldsymbol{v})$ and $\boldsymbol{\psi}^{\mathrm{d}}_{\widehat{K}_3}(\boldsymbol{v})$. (iv) Compute $\int_{\widehat{K}_i} p\nabla\cdot\boldsymbol{u}\,\mathrm{d}\widehat{x}$ for $i\in\{2,3\}$. (*Hint*: use Steps (i) and (iii), and $\int_{\widehat{K}_i} q\nabla\cdot\boldsymbol{v}\,\mathrm{d}\widehat{x} = \int_{\widehat{K}_1}\psi^{\mathrm{g}}_{K_i}(q)\nabla\cdot(\boldsymbol{\psi}^{\mathrm{d}}_{K_i}(\boldsymbol{v}))\,\mathrm{d}\widehat{x}$ (see Exercise 14.3(i)).) (v) Write the linear system corresponding to the statement $(\widehat{B}(\boldsymbol{u}),p)_{L^2(\widehat{U})} := \int_{\widehat{U}} p\nabla\cdot\boldsymbol{u}\,\mathrm{d}\widehat{x} = 0$ for all $\boldsymbol{u}\in\boldsymbol{P}^{\mathrm{g}}_{2,0}(\widehat{U})$, and compute $\mathrm{im}(\widehat{B})^{\perp}$.

Exercise 55.6 (Macroelement partition). Reprove Corollary 55.3 without invoking the partition lemma (Lemma 55.1). (*Hint*: see Brezzi and Bathe [7, Prop.4.2].)

Exercise 55.7 (Macroelement, continuous pressure). Let the assumptions of Proposition 55.5 hold true. (i) Show that there are $c_1,c_2>0$ s.t.

$$\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}}\frac{|b(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \ge c_1\beta_D\|q_h\|_Q - c_2\Big(\sum_{U\in\mathcal{U}_h} h_U^2|q_h|^2_{H^1(U)}\Big)^{\frac12},$$

for all $q_h\in Q_h$ and all $h\in\mathcal{H}$. (*Hint:* use the quasi-interpolation operator $\boldsymbol{\mathcal{I}}^{\mathrm{av}}_{h0}$ and proceed as in the proof of Lemma 54.3.) (ii) Setting $\overline{q}_{hU} := \frac{1}{|U|}\int_U q_h\,\mathrm{d}x$, show that there is c s.t. $|q_{h|U}|_{H^1(U)}\le c\|q_h-\overline{q}_{hU}\|_{L^2(\widehat{U})}$ for all $U\in\mathcal{U}_h$ and all $h\in\mathcal{H}$. (*Hint*: use Lemma 11.7 and the affine geometric mapping $\boldsymbol{T}_U:\widehat{U}\to U$.) (iii) Prove Corollary 55.5. (*Hint*: use Remark 55.4. See also Brezzi and Bathe [7, Prop 4.1].)

Solution to exercises

Exercise 55.1 (Local mass balance). Let $q\in\mathbb{P}_{k,d}$ and $K\in\mathcal{T}_h$. Let us define $q_h\in P^{\mathrm{b}}_k(\mathcal{T}_h)$ by setting $q_{h|K} := q\circ\boldsymbol{T}_K^{-1}$ and $q_{h|K'}=0$ if $K'\neq K$. Let $\overline{q}_h := \frac{1}{|D|}\int_D q_h\,\mathrm{d}x$. This gives $q_h-\overline{q}_h\in P^{\mathrm{b}}_{k,*}(\mathcal{T}_h)$, so that by assumption we have

$$\int_D (q_h-\overline{q}_h)\nabla\cdot\boldsymbol{u}_h\,\mathrm{d}x = \int_D (q_h-\overline{q}_h)g\,\mathrm{d}x.$$

But the compatibility condition $\int_D g\,\mathrm{d}x=0$ and the homogeneous Dirichlet condition enforced on $\boldsymbol{u}_h$ imply that

$$\int_D \overline{q}_h\nabla\cdot\boldsymbol{u}_h\,\mathrm{d}x = \overline{q}_h\int_D\nabla\cdot\boldsymbol{u}_h\,\mathrm{d}x = 0 = \overline{q}_h\int_D g\,\mathrm{d}x = \int_D \overline{q}_h g\,\mathrm{d}x.$$

Hence, we have

$$\int_K (q\circ \boldsymbol{T}_K^{-1})\nabla\cdot\boldsymbol{u}_h\,\mathrm{d}x = \int_D q_h\nabla\cdot\boldsymbol{u}_h\,\mathrm{d}x = \int_D q_h g\,\mathrm{d}x = \int_K (q\circ\boldsymbol{T}_K^{-1})g\,\mathrm{d}x,$$

which proves the assertion.

Exercise 55.2 ($(\mathbb{P}_2,\mathbb{P}_0^{\mathrm{b}})$). Let us verify that the assumptions (i)–(iii) from Lemma 54.2 are met with the operators $\boldsymbol{\Pi}_{1h},\boldsymbol{\Pi}_{2h}$ defined in the proof of Lemma 55.8. Recall that this operators map from $\boldsymbol{V}:=\boldsymbol{W}_0^{1,p}(D)$ to $\boldsymbol{V}_{h0}:=\boldsymbol{P}_{2,0}^{\mathrm{g}}(\mathcal{T}_h)$. The operator $\boldsymbol{\Pi}_{2h}$ is linear, so that the assumption (i) is met. Let us show that it is also the case for the assumption (ii). Let $F\in\mathcal{F}_h^\circ$. Since $\boldsymbol{\Pi}_{2h}(\boldsymbol{v})_{|F}$ is quadratic on F, we can apply Simpson's quadrature rule to infer that

$$\begin{aligned}\int_F \boldsymbol{\Pi}_{2h}(\boldsymbol{v})\,\mathrm{d}s &= \frac{|F|}{6}\Big(\boldsymbol{\Pi}_{2h}(\boldsymbol{v})(\boldsymbol{z}_{1,F}) + 4\boldsymbol{\Pi}_{2h}(\boldsymbol{v})(\boldsymbol{m}_F) + \boldsymbol{\Pi}_{2h}(\boldsymbol{v})(\boldsymbol{z}_{2,F})\Big)\\ &= \frac{2|F|}{3}\boldsymbol{\Pi}_{2h}(\boldsymbol{v})(\boldsymbol{m}_F) = \int_F \boldsymbol{v}\,\mathrm{d}s,\end{aligned}$$

where $\{\boldsymbol{z}_{1,F},\boldsymbol{z}_{2,F}\}$ are the two endpoints of F and $\boldsymbol{m}_F$ is the barycenter of F. Consider now $K\in\mathcal{T}_h$ and let $\mathcal{F}_K^\circ$ be the collection of the mesh interfaces that are faces of K. The above identity implies that

$$\int_K \nabla\cdot(\boldsymbol{v}-\boldsymbol{\Pi}_{2h}(\boldsymbol{v}))\,\mathrm{d}x = \sum_{F\in\mathcal{F}_K^\circ}\int_F (\boldsymbol{v}-\boldsymbol{\Pi}_{2h}(\boldsymbol{v}))\cdot\boldsymbol{n}_K\,\mathrm{d}s = 0,$$

where $\boldsymbol{n}_K$ is the unit outward normal to K. Since Q_h is composed of piecewise constant pressures, we infer that for all $q_h\in Q_h$,

$$b(\boldsymbol{v}-\boldsymbol{\Pi}_{2h}(\boldsymbol{v}),q_h) = \sum_{K\in\mathcal{T}_h} q_{h|K}\int_K \nabla\cdot(\boldsymbol{v}-\boldsymbol{\Pi}_{2h}(\boldsymbol{v}))\,\mathrm{d}x = 0.$$

Hence, the assumption (ii) from Lemma 54.2 also holds true. Let us now turn our attention to the assumption (iii). We define the real numbers

$$c_{1h} := \sup_{\boldsymbol{v}\in\boldsymbol{V}}\frac{\|\boldsymbol{\Pi}_{1h}(\boldsymbol{v})\|_{\boldsymbol{V}}}{\|\boldsymbol{v}\|_{\boldsymbol{V}}},\qquad c_{2h} := \sup_{\boldsymbol{v}\in\boldsymbol{V}}\frac{\|\boldsymbol{\Pi}_{2h}(\boldsymbol{v}-\boldsymbol{\Pi}_{1h}(\boldsymbol{v}))\|_{\boldsymbol{V}}}{\|\boldsymbol{v}\|_{\boldsymbol{V}}}.$$

Owing to the $W_0^{1,p}$-stability of $\mathcal{I}_{h0}^{\mathrm{av}}$ (see Theorem 22.14) and recalling that $\boldsymbol{\Pi}_{1h}:=\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$, we infer that c_{1h} is bounded uniformly w.r.t. $h\in\mathcal{H}$. Moreover, the inverse inequality from Lemma 12.1 (with $r:=p$, $l:=1$, $m:=0$) together with Proposition 12.5, the regularity of the mesh sequence, Hölder's inequality, and the multiplicative trace inequality (12.16) implies that for all $K\in\mathcal{T}_h$ (the value of c changes at each occurrence),

$$\begin{aligned}|\boldsymbol{\Pi}_{2h}(\boldsymbol{v})|_{\boldsymbol{W}^{1,p}(K)} &\le c\,h_K^{-1}|K|^{\frac1p}\sum_{F\in\mathcal{F}_K^\circ}\|\boldsymbol{\Pi}_{2h}(\boldsymbol{v})(\boldsymbol{m}_F)\|_{\ell^2}\\ &\le c\,h_K^{\frac2p-2}\sum_{F\in\mathcal{F}_K^\circ}\|\boldsymbol{v}\|_{\boldsymbol{L}^1(F)} \le c\,h_K^{\frac2p-2}\sum_{F\in\mathcal{F}_K^\circ}|F|^{\frac1{p'}}\|\boldsymbol{v}\|_{\boldsymbol{L}^p(F)}\\ &\le c\,h_K^{\frac1p-1}\sum_{F\in\mathcal{F}_K^\circ}\|\boldsymbol{v}\|_{\boldsymbol{L}^p(F)} \le c\left(h_K^{-1}\|\boldsymbol{v}\|_{\boldsymbol{L}^p(K)} + |\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(K)}\right).\end{aligned}$$

This bound combined with the approximation properties of $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$ (see Theorem 22.14) yields

$$\begin{aligned}|\boldsymbol{\Pi}_{2h}(\boldsymbol{v}-\boldsymbol{\Pi}_{1h}(\boldsymbol{v}))|_{\boldsymbol{W}^{1,p}(K)} &\le c\left(h_K^{-1}\|\boldsymbol{v}-\boldsymbol{\Pi}_{1h}(\boldsymbol{v})\|_{\boldsymbol{L}^p(K)}+|\boldsymbol{v}-\boldsymbol{\Pi}_{1h}(\boldsymbol{v})|_{\boldsymbol{W}^{1,p}(K)}\right)\\ &\le c'\,|\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D_K)},\end{aligned}$$

where D_K is the set of the points composing the mesh cells that have a nonempty intersection with K. Owing to the regularity of the mesh sequence, we conclude by summing over the mesh cells that $|\boldsymbol{\Pi}_{2h}(\boldsymbol{v}-\boldsymbol{\Pi}_{1h}(\boldsymbol{v}))|_{\boldsymbol{W}^{1,p}(D)} \le c|\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}$, i.e., c_{2h} is also uniformly bounded w.r.t. $h\in\mathcal{H}$. This shows that the assumption (iii) from Lemma 54.2 is also met, and this completes the proof.

Exercise 55.3 ($(\mathbb{Q}_k,\mathbb{Q}_{k-1}^{\mathrm{b}})$). Let D be a rectangle and let $\mathcal{T}_h$ be a uniform Cartesian mesh of D. Let $h\in\mathcal{H}$ be the meshsize. Let $\widehat{K}:=(0,1)^2$ be the reference square. We assume that all the geometric transformations $\boldsymbol{T}_K:\widehat{K}\to K\in\mathcal{T}_h$ are homotheties, i.e., using the conventions defined in Table 21.1, we have $\boldsymbol{T}_K(\widehat{\boldsymbol{x}})=\boldsymbol{z}_{1,K}+h\widehat{\boldsymbol{x}}$ for all $\widehat{\boldsymbol{x}}\in\widehat{K}$, where $\boldsymbol{z}_{1,K}$ is the bottom left vertex of K.
(i) Let $\widehat{p}(\widehat{\boldsymbol{x}}):=4(\widehat{x}-\frac{1}{2})(\widehat{y}-\frac{1}{2})$. Note that $\widehat{p}$ takes the alternating values ±1 at the four vertices of $\widehat{K}$. Let p_h be the $\mathbb{Q}_1$-discontinuous pressure field s.t. $p_{h|K}:=\widehat{p}\circ\boldsymbol{T}_K^{-1}=(\psi_K^{\mathrm{g}})^{-1}(\widehat{p})$ for all $K\in\mathcal{T}_h$. We are going to show that p_h is a spurious pressure mode. Let $\boldsymbol{v}_h$ be a continuous $\mathbb{Q}_2$ velocity field with zero trace on ∂D. We have $\int_K(\nabla\cdot\boldsymbol{v}_h)p_h\,\mathrm{d}x=\int_{\widehat{K}}(\nabla\cdot\widehat{\boldsymbol{v}})\widehat{p}\,\mathrm{d}\widehat{x}$, where $\widehat{\boldsymbol{v}}:=\boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{v}_h)$ and $\boldsymbol{\psi}_K^{\mathrm{d}}$ is the contravariant Piola transformation. Since the function $(\nabla\cdot\widehat{\boldsymbol{v}})\widehat{p}$ is a polynomial in $\mathbb{Q}_3$, we can apply the tensor-product version of Simpson's rule to obtain

$$\int_K(\nabla\cdot\boldsymbol{v}_h)p_h\,\mathrm{d}x=\frac{h^2}{36}\sum_{l\in\{1:4\}}\nabla\cdot\boldsymbol{v}_{h|K}(\boldsymbol{z}_{l,K})p_{h|K}(\boldsymbol{z}_{l,K}),$$

where $\boldsymbol{z}_{1,K},\dots,\boldsymbol{z}_{4,K}$ are the four vertices of K. (Recall that all the cells have the same surface h^2 since we assumed that the mesh is uniform.) Let now $\boldsymbol{z}$ be an internal vertex in the mesh and let $K_1,\dots,K_4$ be the four cells sharing $\boldsymbol{z}$. Assume that the cells are enumerated counter-clockwise around $\boldsymbol{z}$ and that K_1 is the top right cell. We infer that $\boldsymbol{T}_{K_i}(\widehat{\boldsymbol{z}}_i)=\boldsymbol{z}$ for all $i\in\{1{:}4\}$. The definition of p_h implies that $p_{h|K_m}(\boldsymbol{z})=(-1)^m$, so that reasoning as in Exercise 54.3(ii) gives

$$\sum_{m\in\{1:4\}}\nabla\cdot\boldsymbol{v}_{h|K_m}(\boldsymbol{z})p_{h|K_m}(\boldsymbol{z})=\sum_{m\in\{1:4\}}(-1)^m\nabla\cdot\boldsymbol{v}_{h|K_m}(\boldsymbol{z})=0.$$

If $\boldsymbol{z}$ is a boundary vertex, but not a corner, a similar argument shows that

$$\sum_{m\in\{1,2\}}(-1)^m\nabla\cdot\boldsymbol{v}_{h|K_m}(\boldsymbol{z})=0,$$

where K_1,K_2 are the two cells sharing $\boldsymbol{z}$, and again $p_{h|K_m}(\boldsymbol{z})=\pm(-1)^m$. Moreover, it is clear that $\nabla\cdot\boldsymbol{v}_h(\boldsymbol{z})=0$ if $\boldsymbol{z}$ is a corner vertex since $\boldsymbol{v}_{h|\partial D}=\boldsymbol{0}$. Finally, using

$$\begin{aligned}\int_D(\nabla\cdot\boldsymbol{v}_h)p_h\,\mathrm{d}x&=\frac{h^2}{36}\sum_{K\in\mathcal{T}_h}\sum_{l\in\{1:4\}}\nabla\cdot\boldsymbol{v}_{h|K}(\boldsymbol{z}_{l,K})p_{h|K}(\boldsymbol{z}_{l,K})\\&=\frac{h^2}{36}\sum_{\boldsymbol{z}\in\mathcal{V}_h}\sum_{m\in\{1:m_{\boldsymbol{z}}\}}\nabla\cdot\boldsymbol{v}_{h|K_m}(\boldsymbol{z})p_{h|K_m}(\boldsymbol{z})\\&=0,\end{aligned}$$

where $m_{\boldsymbol{z}} \in \{1,2,4\}$ is the number of mesh cells sharing $\boldsymbol{z}$, we conclude that p_h is a spurious pressure mode.
(ii) If k is even, the spurious pressure mode p_h is a $\mathbb{Q}_{k-1}$-discontinuous field s.t. $p_{h|K} = (\psi_K^{\mathrm{g}})^{-1}(\widehat{p})$ for all $K \in \mathcal{T}_h$, where

$$\widehat{p}(\widehat{\boldsymbol{x}}) := \prod_{i\in\{1:k-1\}} (\widehat{x} - \widehat{x}_i) \prod_{j\in\{1:k-1\}} (\widehat{y} - \widehat{x}_j),$$

where $(\widehat{x}_1, \dots, \widehat{x}_{k-1})$ are the interior Gauss–Lobatto nodes for the quadrature over $(0,1)$ that is exact for the polynomials of degree $(2k-1)$. Whenever k is odd, we first define a $\mathbb{Q}_{k-1}$ reference field $\widehat{p}$ as above using the interior Gauss–Lobatto nodes. Then we enumerate the mesh cells with two indices as $K_{i,j}$, $i \in \mathcal{I}, j \in \mathcal{J}$, where $\boldsymbol{z}_{1,K_{i,j}} := (ih, jh)$, and we define the spurious pressure field by setting $p_{h|K_{i,j}} := (-1)^{i+j}\widehat{p} \circ \boldsymbol{T}_{K_{i,j}}$. This way, we still have that for every interior mesh vertex $\boldsymbol{z}$ shared by the cells $K_1, \dots, K_4$, $p_{h|K_m}(\boldsymbol{z}) = (-1)^m c_k$, where $c_k = \prod_{i\in\{1:k-1\}} \widehat{x}_i^2$, with the same modifications as above when the mesh vertex $\boldsymbol{z}$ lies at the boundary.

Exercise 55.4 ($(\mathbb{P}_1^{\mathrm{CR}}, \mathbb{P}_0^{\mathrm{b}})$). It is straightforward to adapt the error estimate from Theorem 53.17 to the present nonconforming setting by proceeding as in §36.3 to handle the discrete bilinear form a_h.

Exercise 55.5 ($(\mathbb{P}_2, \mathbb{P}_1^{\mathrm{b}})$, HCT mesh). (i) The $\mathbb{P}_1$ shape functions on $\widehat{K}_1$ associated with the vertices $\widehat{\boldsymbol{z}}_1$, $\widehat{\boldsymbol{z}}_2$, and $\widehat{\boldsymbol{z}}_4$ are, respectively,

$$\theta_1(x,y) = 1 - x - 2y, \quad \theta_2(x,y) = x - y, \quad \theta_4(x,y) = 3y.$$

The $\mathbb{P}_2$ shape functions on $\widehat{K}_1$ associated with the nodes $\widehat{\boldsymbol{m}}_{14}$, $\widehat{\boldsymbol{m}}_{24}$, and $\widehat{\boldsymbol{z}}_4$ are, respectively,

$$\psi_7(x,y) = 12y(1-x-2y), \quad \psi_8(x,y) = 12y(x-y), \quad \psi_{10}(x,y) = 3y(6y-1).$$

We have $p_{|\widehat{K}_1} = p_1\theta_1 + p_2\theta_2 + p_3\theta_3$ and $\boldsymbol{u}_{|\widehat{K}_1} = (u_7, v_7)^{\mathsf{T}}\psi_7 + (u_8, v_8)^{\mathsf{T}}\psi_8 + (u_{10}, v_{10})^{\mathsf{T}}\psi_{10}$. The rest of the computation can be done by hand or by using any symbolic comping software:

$$\begin{aligned}\int_{\widehat{K}_1} p\nabla\cdot\boldsymbol{u}\,\mathrm{d}\widehat{x} = {}& (-u_7 + u_8 + 4v_7 + 2v_8)p_1 \\ & + (-u_7 + u_8 + v_7 + 5v_8)p_2 \\ & + (-2u_7 + 2u_8 - v_7 + v_8 + 3v_{10})p_3.\end{aligned}$$

(ii) We have $\boldsymbol{T}_{\widehat{K}_2}(\widehat{\boldsymbol{z}}_1) = \widehat{\boldsymbol{z}}_2$, $\boldsymbol{T}_{\widehat{K}_2}(\widehat{\boldsymbol{z}}_2) = \widehat{\boldsymbol{z}}_3$, $\boldsymbol{T}_{\widehat{K}_2}(\widehat{\boldsymbol{z}}_4) = \widehat{\boldsymbol{z}}_4$, and $\boldsymbol{T}_{\widehat{K}_3}(\widehat{\boldsymbol{z}}_1) = \widehat{\boldsymbol{z}}_3$, $\boldsymbol{T}_{\widehat{K}_3}(\widehat{\boldsymbol{z}}_2) = \widehat{\boldsymbol{z}}_1$, $\boldsymbol{T}_{\widehat{K}_3}(\widehat{\boldsymbol{z}}_4) = \widehat{\boldsymbol{z}}_4$.
(iii) Using the expressions for $\mathbb{J}_{\widehat{K}_2}$ and $\mathbb{J}_{\widehat{K}_3}$, we obtain

$$\begin{aligned}\boldsymbol{\psi}_{\widehat{K}_2}^{\mathrm{d}}(\boldsymbol{v}) &= \det(\mathbb{J}_{\widehat{K}_2})\mathbb{J}_{\widehat{K}_2}^{-1}\begin{pmatrix}u\\v\end{pmatrix} = \begin{pmatrix}v\\-u-v\end{pmatrix},\\ \boldsymbol{\psi}_{\widehat{K}_3}^{\mathrm{d}}(\boldsymbol{v}) &= \det(\mathbb{J}_{\widehat{K}_3})\mathbb{J}_{\widehat{K}_3}^{-1}\begin{pmatrix}u\\v\end{pmatrix} = \begin{pmatrix}-u-v\\u\end{pmatrix}.\end{aligned}$$

(iv) Using the above results and the hint, we make the change of variables $p \to q$, $7 \to 8$, $8 \to 9$,

$u \to v$, $v \to -u - v$, and obtain

$$\begin{aligned}\int_{\widehat{K}_2} p\nabla\cdot\boldsymbol{u}\,\mathrm{d}\widehat{x} &= \int_{\widehat{K}_1} \psi^{\mathrm{g}}_{\widehat{K}_2}(p)\nabla\cdot(\boldsymbol{\psi}^{\mathrm{d}}_{\widehat{K}_2}(\boldsymbol{u}))\,\mathrm{d}\widehat{x}\\ &= (-v_8+v_9+4(-u_8-v_8)+2(-u_9-v_9)q_1\\ &\quad + (-v_8+v_9+(-u_8-v_8)+5(-u_9-v_9))q_2\\ &\quad + (-2v_8+2v_9-(-u_8-v_8)+(-u_9-v_9)+3(-u_{10}-v_{10}))q_3\\ &= (-4u_8-2u_9-5v_8-v_9)q_1\\ &\quad + (-u_8-5u_9-2v_8-4v_9)q_2\\ &\quad + (u_8-u_9-3u_{10}-v_8+v_9-3v_{10})q_3.\end{aligned}$$

Similarly, making the change of variables $p \to s$, $7 \to 9$, $8 \to 7$, $u \to -u - v$, $v \to u$, we obtain

$$\begin{aligned}\int_{\widehat{K}_3} p\nabla\cdot\boldsymbol{u}\,\mathrm{d}\widehat{x} &= \int_{\widehat{K}_1} \psi^{\mathrm{g}}_{\widehat{K}_3}(p)\nabla\cdot(\boldsymbol{\psi}^{\mathrm{d}}_{\widehat{K}_3}(\boldsymbol{u}))\,\mathrm{d}\widehat{x}\\ &= (u_9+v_9+(-u_7-v_7)+4u_9+2u_7)s_1\\ &\quad + ((u_9+v_9)+(-u_7-v_7)+u_9+5u_7)s_2\\ &\quad + ((2u_9+2v_9)+(-2u_7-2v_7)-u_9+u_7+3u_{10})s_3\\ &= (u_7+5u_9-v_7+v_9)s_1\\ &\quad + (4u_7+2u_9-v_7+v_9)s_2\\ &\quad + (-u_7+u_9+3u_{10}-2v_7+2v_9)s_3.\end{aligned}$$

(v) The identity $(\widehat{B}(\boldsymbol{u}),p)_{L^2(\widehat{U})} = 0$ for all $\boldsymbol{u} \in \boldsymbol{P}^{\mathrm{g}}_{2,0}(\widehat{U})$ is equivalent to $\mathsf{U}^\mathsf{T}\mathcal{B}\mathsf{P} = 0$ with the vectors

$$\begin{aligned}\mathsf{U} &:= (u_7,u_8,u_9,u_{10},v_7,v_8,v_9,v_{10})^\mathsf{T} \in \mathbb{R}^8,\\ \mathsf{P} &:= (p_1,p_2,p_3,q_1,q_2,q_3,s_1,s_2,s_3)^\mathsf{T} \in \mathbb{R}^9,\end{aligned}$$

and the matrix $\mathcal{B} \in \mathbb{R}^{8\times 9}$ s.t.

$$\mathcal{B} := \left[\begin{array}{ccc:ccc:ccc} -1&-1&-2&0&0&0&1&4&-1\\ 1&1&2&-4&-1&1&0&0&0\\ 0&0&0&-2&-5&-1&5&2&1\\ 0&0&0&0&0&-3&0&0&3\\ \hdashline 4&1&-1&0&0&0&-1&-1&-2\\ 2&5&1&-5&-2&-1&0&0&0\\ 0&0&0&-1&-4&1&1&1&2\\ 0&0&3&0&0&-3&0&0&0 \end{array}\right].$$

One can verify that the matrix $\mathcal{B}$ has full row rank and that $\ker(\mathcal{B}) = \mathrm{span}((1,\dots,1)^\mathsf{T})$. Hence $\mathrm{im}(\widehat{B})^\perp = \mathrm{span}(\mathbb{1}_{\widehat{U}})$.

Exercise 55.6 (Macroelement partition). Let $q_h \in Q_h$. For all $U \in \mathcal{U}_h$, we set $\overline{q}_U := \frac{1}{|U|}\int_U q_h\,\mathrm{d}x$ and $\overline{q}_h := \sum_{U\in\mathcal{U}} \overline{q}_U \mathbb{1}_U$. Proceeding as in the proof of Corollary 55.3, we infer that

$$\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \ge \sup_{\boldsymbol{v}_h\in\boldsymbol{V}^1_{h0}} \frac{|b(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \ge \beta_{1h}\|q_h-\overline{q}_h\|_Q,$$

with $\boldsymbol{V}_h^1 := \sum_{U\in\mathcal{U}_h} \boldsymbol{V}_{h0}(U)$. Moreover, we have

$$\begin{aligned}\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} &\geq \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h,\overline{q}_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} - \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h,q_h-\overline{q}_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}}\\ &\geq \beta_{2h}\|\overline{q}_h\|_Q - \|q_h-\overline{q}_h\|_Q,\end{aligned}$$

since $\|\nabla\cdot\boldsymbol{v}_h\|_{L^2(D)} \leq \|\boldsymbol{v}\|_{\boldsymbol{V}} = |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}$. We infer that

$$\Big(\frac{1}{\beta_{1h}}+1\Big)\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \geq \beta_{2h}\|\overline{q}_h\|_Q.$$

In conclusion, we obtain

$$\left(\frac{1}{\beta_{1h}}+\frac{1}{\beta_{2h}}\Big(\frac{1}{\beta_{1h}}+1\Big)\right)\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \geq \|\overline{q}_h\|_Q + \|q_h-\overline{q}_h\|_Q \geq \|q_h\|_Q.$$

This proves the assertion.

Exercise 55.7 (Macroelement, continuous pressure). (i) Owing to the inf-sup condition (53.9), there exists $\beta_D > 0$ such that for all $q_h \in Q_h$, there is $\boldsymbol{v}(q_h) \in \boldsymbol{V} := \boldsymbol{H}_0^1(D)$ such that $\frac{|b(\boldsymbol{v}(q_h),q_h)|}{\|\boldsymbol{v}(q_h)\|_{\boldsymbol{V}}} \geq \beta_D\|q_h\|_Q$. Let $q_h \in Q_h$. Recalling that $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$ is the $\mathbb{R}^d$-valued version of the H_0^1-conforming quasi-interpolation operator introduced in §22.4.2, we have

$$\begin{aligned}\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} &\geq \frac{|b(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}(q_h)),q_h)|}{\|\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}(q_h))\|_{\boldsymbol{V}}}\\ &\geq c\,\frac{|b(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}(q_h)),q_h)|}{\|\boldsymbol{v}(q_h)\|_{\boldsymbol{V}}}\\ &\geq c\,\frac{|b(\boldsymbol{v}(q_h),q_h)|}{\|\boldsymbol{v}(q_h)\|_{\boldsymbol{V}}} - c\,\frac{|b(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}(q_h))-\boldsymbol{v}(q_h),q_h)|}{\|\boldsymbol{v}(q_h)\|_{\boldsymbol{V}}}\\ &\geq c\,\beta_D\|q_h\|_Q - c\,\frac{|b(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}(q_h))-\boldsymbol{v}(q_h),q_h)|}{\|\boldsymbol{v}(q_h)\|_{\boldsymbol{V}}}.\end{aligned}$$

Using that Q_h is H^1-conforming and since we are enforcing the homogeneous Dirichlet boundary condition on the velocity over the entire boundary ∂D, one integration by parts together with the Cauchy–Schwarz inequality and the approximation properties of $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$ gives

$$\begin{aligned}|b(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}(q_h))-\boldsymbol{v}(q_h),q_h)| &= \Big|\sum_{K\in\mathcal{T}_h}\int_K (\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}(q_h))-\boldsymbol{v}(q_h))\cdot\nabla q_h\,\mathrm{d}x\Big|\\ &\leq c\sum_{K\in\mathcal{T}_h} h_K|\boldsymbol{v}(q_h)|_{\boldsymbol{H}^1(K)}|q_h|_{H^1(K)}\\ &\leq c\,\|\boldsymbol{v}(q_h)\|_{\boldsymbol{V}}\Big(\sum_{K\in\mathcal{T}_h} h_K^2|q_h|_{H^1(K)}^2\Big)^{\frac12}.\end{aligned}$$

Since $h_K \leq h_U$ for all $K \in U$, we have

$$\frac{|b(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}(q_h))-\boldsymbol{v}(q_h),q_h)|}{\|\boldsymbol{v}(q_h)\|_{\boldsymbol{V}}} \leq c\Big(\sum_{U\in\mathcal{U}_h} h_U^2|q_h|_{H^1(U)}^2\Big)^{\frac12}.$$

(ii) Denoting $\overline{q}_{hU} := \frac{1}{|U|}\int_U q_h \,\mathrm{d}x$, using the affine geometric mapping $\boldsymbol{T}_U : \widehat{U} \to U$, and invoking the shape-regularity of the macroelement partition, we have

$$h_U|q_h|_{H^1(U)} = h_U|q_{h|U} - \overline{q}_{hU}|_{H^1(U)} \le c\, h_U \|\mathbb{J}_U^{-1}\| |\det(\mathbb{J}_U)|^{\frac{1}{2}} |\widehat{q}_h - \overline{q}_{hU}|_{H^1(\widehat{U})},$$

where $\widehat{q}_h := q_{h|U} \circ \boldsymbol{T}_U$ (see Lemma 11.7). Since $\inf_{h\in\mathcal{H}} \max_{U\in\mathcal{U}_h} \mathrm{card}\{K \subset U\} < \infty$ implies that $\mathrm{span}\{q_{h|U} \circ \boldsymbol{T}_U\}$ is a finite-dimensional space, we invoke the equivalence of norms and infer that

$$h_U|q_{h|U}|_{H^1(U)} \le c\, h_U \|\mathbb{J}_U^{-1}\| |\det(\mathbb{J}_U)|^{\frac{1}{2}} \|\widehat{q}_h - \overline{q}_{hU}\|_{L^2(\widehat{U})},$$

where c is uniform w.r.t. $h \in \mathcal{H}$ (because the dimension of $\mathrm{span}\{q_{h|U} \circ \boldsymbol{T}_U\}$ is bounded from above uniformly w.r.t. $h \in \mathcal{H}$). Invoking again Lemma 11.7, we obtain

$$h_U|q_{h|U}|_{H^1(U)} \le c\, h_U \|\mathbb{J}_U^{-1}\| \|q_h - \overline{q}_{hU}\|_{L^2(U)}.$$

Since the geometric mapping $\boldsymbol{T}_U$ is affine, we have $h_U\|\mathbb{J}_U^{-1}\| \le c$ (see (11.3)), and we conclude that for all $U \in \mathcal{U}_h$,

$$|q_{h|U}|_{H^1(U)} \le c\, \|q_h - \overline{q}_{hU}\|_{L^2(U)}.$$

(iii) Combining the results of Steps (i) and (ii) shows that

$$c_1 \Big(\sum_{U\in\mathcal{U}_h} \|q_{h|U} - \overline{q}_{hU}\|^2_{L^2(U)} \Big)^{\frac{1}{2}} + \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \ge c_2\beta_D \|q_h\|_Q.$$

Now, we invoke Remark 55.4 and observe that

$$\sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \ge \sup_{\boldsymbol{v}_h \in \boldsymbol{V}^1_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \ge \beta_{1h} \Big(\sum_{U\in\mathcal{U}_h} \|q_{h|U} - \overline{q}_{hU}\|^2_{L^2(U)} \Big)^{\frac{1}{2}},$$

with $\boldsymbol{V}^1_{h0} := \sum_{U\in\mathcal{U}_h} \boldsymbol{V}_{h0}(U)$. Combining the above two bounds shows that

$$\Big(\frac{c_1}{\beta_{1h}} + 1 \Big) \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \ge c_2\beta_D \|q_h\|_Q.$$

In conclusion, the inf-sup condition (55.1) holds uniformly w.r.t. $h \in \mathcal{H}$ if

$$\inf_{h\in\mathcal{H}} \beta_{1h} := \inf_{h\in\mathcal{H}} \min_{U\in\mathcal{U}_h} \beta_{1h} > 0.$$

Chapter 56

Friedrichs' systems

Exercises

Exercise 56.1 (Robin condition). Show how to enforce the Robin boundary condition $\gamma u - \boldsymbol{\sigma}\cdot\boldsymbol{n} = 0$ on ∂D (with $\gamma \in L^\infty(\partial D)$ and $\gamma \geq 0$ a.e. on ∂D) in the framework of §56.2.2.

Exercise 56.2 (Linear elasticity). Consider the linear elasticity model from §42.1. Verify that $\mathbb{s} - \frac{1}{d+\theta}\operatorname{tr}(\mathbb{s})\mathbb{I}_d = \mu(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^\mathsf{T})$ with $\theta := \frac{2\mu}{\lambda}$ and that $\frac{1}{2}\nabla\cdot(\mathbb{s} + \mathbb{s}^\mathsf{T}) + \boldsymbol{f} = \boldsymbol{0}$. Write this system using Friedrichs' formalism. (*Hint*: identify $\mathbb{s} \in \mathbb{R}^{d\times d}$ with a vector $\boldsymbol{s} \in \mathbb{R}^{d^2}$ by setting $\boldsymbol{s}_{[ij]} := \mathbb{s}_{ij}$ with $[ij] := d(j-1)+i$ for all $i,j \in \{1{:}d\}$.) Verify (56.1a)-(56.1b) and that the upper left block of $\mathcal{K}$, say $\mathcal{K}^{ss}$, is positive definite. What happens in the incompressible limit $\lambda \to \infty$?

Exercise 56.3 (Positivity, locality). (i) Reprove Theorem 56.9 by replacing the assumption made on $\mathcal{K}$ by those stated in Remark 56.12. (ii) Let $D := (0,a)\times(-1,1)$, $a > 0$, and let $K : L^2(D) \to L^2(D)$ be such that $K(v)(x,y) := v(x,y) - \frac{\sigma}{2}\int_{-1}^{+1} v(x,\xi)\,\mathrm{d}\xi$ with $\sigma \in [0,1)$. Assuming $\mathcal{X} := 0$, prove that K satisfies the assumptions from Remark 56.12.

Exercise 56.4 (Wave equation). Consider the wave equation $\frac{\partial^2 v}{\partial t^2} - \frac{\partial^2 v}{\partial x^2} = f$ in $D := (0,1)\times(-1,1)$ with the boundary conditions $\frac{\partial v}{\partial t}(t,\pm 1) = 0$ for all $t \in (0,1)$ and $\frac{\partial v}{\partial t}(0,x) = \frac{\partial v}{\partial x}(0,x) = 0$ for all $x \in (-1,1)$. Recast this problem as a Friedrichs' system and identify the boundary fields $\mathcal{N}$ and $\mathcal{M}$. (*Hint*: set $u := e^{-\lambda t}(\frac{\partial v}{\partial t}, \frac{\partial v}{\partial x})$ with $\lambda > 0$.)

Exercise 56.5 (Partial positivity). Assume that there is an orthogonal projection operator $\mathcal{P} \in \mathbb{C}^{m\times m}$ (i.e., $\mathcal{P}^\mathsf{T} = \mathcal{P}$ and $\mathcal{P}^2 = \mathcal{P}$) such that

$$\mathcal{K} + \mathcal{K}^\mathsf{T} - \mathcal{X} \geq 2\mu_0\mathcal{P} \text{ a.e. in } D, \tag{56.1a}$$

$$\sup_{w\in L}\frac{|(A(v),w)_L|}{\|w\|_L} \geq \alpha\|(\mathbb{I}_m - \mathcal{P})(v)\|_L - \lambda\|\mathcal{P}(v)\|_L \text{ for all } v \in V_0, \tag{56.1b}$$

$$\|\mathcal{P}(w)\|_L \geq \gamma\|(\mathbb{I}_m - \mathcal{P})(w)\|_L \text{ for all } w \in \tilde{V}_0 \text{ s.t. } \tilde{A}(w) = 0, \tag{56.1c}$$

with $\mu_0 > 0$ $\alpha > 0$, $\gamma > 0$, λ, and $\tilde{V}_0 := \ker(M^* + N)$. (i) Assume (56.1a), (56.1b), (56.27), and (56.1). Prove that $A : V_0 \to L$ is an isomorphism. (*Hint*: adapt the proof of Theorem 56.9.) (ii) Verify (56.1a) for Darcy's equations with $\mu := 0$ and a Dirichlet boundary condition on p. (*Hint*: use a Poincaré–Steklov inequality.)

Exercise 56.6 ((BNB1) for Darcy and Maxwell). (i) Prove the condition (BNB1) for Darcy's equations with Dirichlet or Neumann condition. (*Hint*: use the test function $w := (\boldsymbol{\tau}, q) := (\boldsymbol{\sigma} + \mathrm{d}\nabla p, p + \mu^{-1}\nabla\cdot\boldsymbol{\sigma})$.) (ii) Do the same for Maxwell's equations with the condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$ or $\boldsymbol{E}\times\boldsymbol{n} = \boldsymbol{0}$. (*Hint*: use the test function $w := (\boldsymbol{e}, \boldsymbol{b}) := (e^{-\mathrm{i}\theta}(\boldsymbol{E} - \mathrm{i}\frac{1}{\sigma}\nabla\times\boldsymbol{H}), e^{\mathrm{i}\theta}(\boldsymbol{H} + \frac{1}{\omega\mu}\nabla\times\boldsymbol{E}))$ where $\theta := \frac{\pi}{4}$.)

Exercise 56.7 (Boundary operator for Darcy and Maxwell). (i) Verify that M defined in (56.35) satisfies (56.27) and that it can be used to enforce a Dirichlet boundary condition on p. (*Hint*: use Theorem 4.15.) How should M be modified to enforce a Neumann condition? (ii) Verify that M defined in (56.36) satisfies (56.27) and that it can be used to enforce the boundary condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$. (*Hint*: use the surjectivity of traces from $H^1(D)$ onto $H^{\frac{1}{2}}(\partial D)$ and (4.11).) How should M be modified to enforce the boundary condition $\boldsymbol{E}\times\boldsymbol{n} = \boldsymbol{0}$?

Exercise 56.8 (Separation assumption). Let $D := \{(x_1, x_2) \in \mathbb{R}^2 \mid 0 < x_2 < 1 \text{ and } |x_1| < x_2\}$ with $\boldsymbol{\beta} := (1, 0)^\mathsf{T}$. Let $V := \{v \in L^2(D) \mid \boldsymbol{\beta}\cdot\nabla v \in L^2(D)\}$. Verify that the function $u(x_1, x_2) := x_2^\alpha$ is in V for $\alpha > -1$, but $u_{|\partial D} \in L^2(|\boldsymbol{\beta}\cdot\boldsymbol{n}|; \partial D)$ only if $\alpha > -\frac{1}{2}$.

Exercise 56.9 (Semi-norm $|\cdot|_M$). Let V be a complex Hilbert space, $N, M \in \mathcal{L}(V; V')$, and let $V_0 := \ker(M - N)$. Assume $N = N^*$ and $\Re(\langle M(v), v\rangle_{V',V}) \geq 0$ for all $v \in V$. Let $|v|_M^2 := \Re(\langle M(v), v\rangle_{V',V})$ for all $v \in V$. Prove that $|\langle N(v), w\rangle_{V',V}| \leq |v|_M|w|_M$ for all $v, w \in V_0$.

Solution to exercises

Exercise 56.1 (Robin condition). We can take

$$\mathcal{M} := \left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \boldsymbol{n} \\ \hdashline -\boldsymbol{n}^\mathsf{T} & 2\gamma \end{array}\right],$$

so that

$$\mathcal{M} - \mathcal{N} = 2\left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \mathbb{O}_{d\times 1} \\ \hdashline -\boldsymbol{n}^\mathsf{T} & \gamma \end{array}\right], \qquad \mathcal{M} + \mathcal{N} = 2\left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \boldsymbol{n} \\ \hdashline \mathbb{O}_{1\times d} & \gamma \end{array}\right].$$

Then (56.7a) holds true since $\gamma \geq 0$ a.e. on ∂D. To see that (56.7b) is satisfied, one can write $(\boldsymbol{\tau}, q) = (\gamma v\boldsymbol{n}, v) + (\boldsymbol{\tau} - \gamma v\boldsymbol{n}, 0)$ and observe that $(\gamma v\boldsymbol{n}, v) \in \ker(\mathcal{M} - \mathcal{N})$ and $(\boldsymbol{\tau} - \gamma v\boldsymbol{n}, 0) \in \ker(\mathcal{M} + \mathcal{N})$.

Exercise 56.2 (Linear elasticity). Taking the trace of $\mathbb{s} - \frac{1}{d+\theta}\operatorname{tr}(\mathbb{s})\mathbb{I}_d = \mu(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^\mathsf{T})$, we infer that $\frac{\theta}{d+\theta}\operatorname{tr}(\mathbb{s}) = 2\mu\nabla\cdot\boldsymbol{u}$, so that $\mathbb{s} = \frac{2\mu}{\theta}(\nabla\cdot\boldsymbol{u})\mathbb{I}_d + \frac{2\mu}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^\mathsf{T})$, which coincides with the constitutive law (42.3) since $\theta = \frac{2\mu}{\lambda}$. Since $\mathbb{s}$ is symmetric, we also infer that $\frac{1}{2}\nabla\cdot(\mathbb{s} + \mathbb{s}^\mathsf{T}) = \nabla\cdot\mathbb{s} = -\boldsymbol{f}$ owing to (42.1). Using the suggested identification between $\mathbb{s} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{s} \in \mathbb{R}^{d^2}$, the above PDEs can be cast into Friedrichs' formalism by setting $m := d^2 + d$ and

$$\mathcal{K} := (2\mu)^{-1}\left[\begin{array}{c:c} \mathcal{K}^{ss} & \mathbb{O}_{d^2\times d} \\ \hdashline \mathbb{O}_{d\times d^2} & \mathbb{O}_{d\times d} \end{array}\right], \qquad \mathcal{A}^k := \left[\begin{array}{c:c} \mathbb{O}_{d^2\times d^2} & \mathcal{E}^k \\ \hdashline (\mathcal{E}^k)^\mathsf{T} & \mathbb{O}_{d\times d} \end{array}\right],\ k \in \{1{:}d\},$$

with $\mathcal{K}^{ss} \in \mathbb{R}^{d^2\times d^2}$ s.t. $\mathcal{K}^{ss}_{[ij][kl]} := \delta_{ik}\delta_{jl} - \frac{1}{d+\theta}\delta_{ij}\delta_{kl}$ and $\mathcal{E}^k \in \mathbb{R}^{d^2\times d}$ s.t. $\mathcal{E}^k_{[ij],l} := -\frac{1}{2}(\delta_{ik}\delta_{jl} + \delta_{il}\delta jk)$, where the δ's are Kronecker symbols. Assumption (56.1a) holds true since all the fields are constant

(so that $\mathcal{X} = \mathbb{O}_{m\times m}$). Assumption (56.1b) holds true since $\mathcal{A}^k$ is symmetric. Finally, we observe that

$$2\mu(\mathcal{K}^{ss}\boldsymbol{s},\boldsymbol{s})_{\ell^2(\mathbb{R}^{d^2})} = \frac{\theta}{d+\theta}\|\mathbb{s}\|^2_{\ell^2(\mathbb{R}^{d\times d})} + \frac{d}{d+\theta}\|\mathbb{s} - d^{-1}\operatorname{tr}(\mathbb{s})\mathbb{I}_d\|^2_{\ell^2(\mathbb{R}^{d\times d})}.$$

In the incompressible limit, $\theta \to 0$ and the full control on $\mathbb{s}$ is lost, and only the control on the deviatoric part $\mathbb{s} - d^{-1}\operatorname{tr}(\mathbb{s})\mathbb{I}_d$ remains.

Exercise 56.3 (Positivity, locality). (i) Defining the formal adjoint by $\tilde{A}(v) = K^*(v) - \mathcal{X}v - A_1(v)$, the results in Lemma 56.8 still hold true owing to the following identity:

$$\begin{aligned}(A(v),v)_L &= \tfrac12(A(v),v)_L + \tfrac12(A(v),v)_L = \tfrac12(A(v),v)_L + \tfrac12(v,\tilde{A}(v))_L + \tfrac12(N(v),v)_{V',V}\\ &= \tfrac12(A(v),v)_L + \tfrac12(v,K^*(v) - \mathcal{X}v - A_1(v))_L + \tfrac12(N(v),v)_{V',V}\\ &= \tfrac12(A(v),v)_L + \tfrac12(v,K^*(v) - \mathcal{X}v + K(v) - A(v))_L + \tfrac12(N(v),v)_{V',V}\\ &= \tfrac12(A(v),v)_L - \tfrac12\overline{(A(v),v)}_L + \tfrac12(v,(K^*+K)(v) - \mathcal{X}v)_L + \tfrac12(N(v),v)_{V',V}\\ &= \tfrac12(A(v),v)_L - \tfrac12\overline{(A(v),v)}_L + \tfrac12((K^*+K)(v) - \mathcal{X}v,v)_L + \tfrac12(N(v),v)_{V',V}.\end{aligned}$$

The rest of the proof of Theorem 56.9 is unchanged.
(ii) Notice that K is a bounded operator on $L^2(D)$ (apply the triangle inequality and the Cauchy–Schwarz inequality). Moreover, we have

$$\begin{aligned}(K(v),w)_{L^2(D)} &= \int_0^a\int_{-1}^{+1} v(x,y)w(x,y)\,\mathrm{d}x\,\mathrm{d}y - \frac{\sigma}{2}\int_0^a\int_{-1}^{+1}\int_{-1}^{+1} v(x,\xi)w(x,y)\,\mathrm{d}\xi\,\mathrm{d}x\,\mathrm{d}y\\ &= \int_0^a\int_{-1}^{+1} v(x,y)w(x,y)\,\mathrm{d}x\,\mathrm{d}y - \frac{\sigma}{2}\int_0^a\int_{-1}^{+1}\int_{-1}^{+1} v(x,y)w(x,\xi)\,\mathrm{d}\xi\,\mathrm{d}x\,\mathrm{d}y\\ &= \int_0^a\int_{-1}^{+1} v(x,y)\left(w(x,y) - \frac{\sigma}{2}\int_{-1}^{+1} w(x,\xi)\,\mathrm{d}\xi\right)\mathrm{d}x\,\mathrm{d}y = (v,K(w))_{L^2(D)}.\end{aligned}$$

Hence, $K = K^*$, i.e., K is self-adjoint. Using the inequality

$$\left|\int_{-1}^{+1} v(x,y)\frac12\int_{-1}^{+1} v(x,\xi)\,\mathrm{d}\xi\,\mathrm{d}y\right| = \frac12\left(\int_{-1}^{+1} v(x,\xi)\,\mathrm{d}\xi\right)^2 \le \int_{-1}^{+1} v(x,\xi)^2\,\mathrm{d}\xi,$$

we infer that

$$\begin{aligned}((K+K^*)(v),v)_{L^2(D)} &= 2(K(v),v)_{L^2(D)}\\ &\ge 2\|v\|^2_{L^2(D)} - 2\sigma\|v\|^2_{L^2(D)} = 2(1-\sigma)\|v\|^2_{L^2(D)},\end{aligned}$$

and this proves the statement with $\mu_0 := 1 - \sigma$.

Exercise 56.4 (Wave equation). We obtain

$$\mathcal{K} := \left[\begin{array}{c:c}\lambda & 0\\ \hdashline 0 & \lambda\end{array}\right], \quad \mathcal{A}^t := \left[\begin{array}{c:c}1 & 0\\ \hdashline 0 & 1\end{array}\right], \quad \mathcal{A}^x := \left[\begin{array}{c:c}0 & -1\\ \hdashline -1 & 0\end{array}\right],$$

and the source term is $(e^{-\lambda t}f, 0)^{\mathsf{T}}$, so that the properties (56.1) hold true. Moreover, we have

$$\mathcal{N} := \left[\begin{array}{c:c}n_t & -n_x\\ \hdashline -n_x & n_t\end{array}\right],$$

where $\boldsymbol{n} := (n_t, n_x)^\mathsf{T}$ is the outward unit normal to D. On the side $\{t = 0, x \in (-1,1)\}$ where $\boldsymbol{n} = (-1,0)^\mathsf{T}$, we can take $\mathcal{M} := -\mathcal{N}$ enforcing the conditions $\frac{\partial v}{\partial t} = \frac{\partial v}{\partial x} = 0$. On the sides $\{x = \pm 1, t \in (0,1)\}$ where $\boldsymbol{n} = (0, \pm 1)^\mathsf{T}$, we can take

$$\mathcal{M} := \pm \left[\begin{array}{c:c} 0 & -1 \\ \hdashline 1 & 0 \end{array}\right]$$

so as to enforce the condition $\frac{\partial v}{\partial t} = 0$. On the side $\{t = 1, x \in (-1,1)\}$ where $\boldsymbol{n} = (1,0)^\mathsf{T}$, we can take $\mathcal{M} := \mathcal{N}$ so that no condition is enforced. We see that the properties (56.7) hold true in all the cases.

Exercise 56.5 (Partial positivity). (i) We outline the differences with respect to the proof of Theorem 56.9. Concerning (BNB1), let $v \in V_0$. Proceeding as before, using (56.1a), and using that $\mathcal{P}$ is an orthogonal projection operator gives $\mu_0 \|\mathcal{P}(v)\|_L^2 \le \Re\big(a(v,v)\big)$. Owing to the triangle inequality and (56.1b), we infer that

$$\begin{aligned}\mu_0\|\mathcal{P}(v)\|_L^2 &\le \frac{\Re((A(v),v)_L)}{\|v\|_L}\|v\|_L \\ &\le \frac{|(A(v),v)_L|}{\|v\|_L}(\|\mathcal{P}(v)\|_L + \|(\mathbb{I}_m - \mathcal{P})(v)\|_L) \\ &\le \sup_{w\in L}\frac{|(A(v),w)_L|}{\|w\|_L}\bigg(\Big(1+\frac{\lambda}{\alpha}\Big)\|\mathcal{P}(v)\|_L + \frac{1}{\alpha}\sup_{w\in L}\frac{|(A(v),w)_L|}{\|w\|_L}\bigg),\end{aligned}$$

yielding

$$\mu_0^{\frac12}\|\mathcal{P}(v)\|_L \le c_1 \sup_{w\in L}\frac{|(A(v),w)_L|}{\|w\|_L},$$

with $c_1 := \Big(\frac{(1+\frac{\lambda}{\alpha})^2}{\mu_0} + \frac{2}{\alpha}\Big)^{\frac12}$. Owing to (56.1b), we infer that

$$\mu_0^{\frac12}\|v\|_L \le c_2 \sup_{w\in L}\frac{|(A(v),w)_L|}{\|w\|_L},$$

with $c_2 := c_1(1+\frac{\lambda}{\alpha}) + \frac{1}{\alpha}\mu_0^{\frac12}$. The rest of the proof of (BNB1) proceeds as before. Concerning (BNB2), let $w \in L$ be such that $a(v,w) = 0$ for all $v \in V_0$. Proceeding as before, we infer successively that $\tilde{A}(w) = 0$, $w \in \tilde{V}_0$, and $0 = \Re\big((\tilde{A}(w),w)_L\big) \ge \mu_0\|\mathcal{P}(w)\|_L^2$, so that $\mathcal{P}(w) = 0$. Invoking (56.1c), we conclude that $w = 0$.

(ii) We set $\mathcal{P}(\boldsymbol{\sigma}, p) := (\boldsymbol{\sigma}, 0)$. Moreover, $V_0 = \tilde{V}_0 = \boldsymbol{H}(\mathrm{div}; D) \times H_0^1(D)$ since a Dirichlet condition is enforced on p, so that we can use the Poincaré–Steklov inequality $C_{\mathrm{PS}}\|p\|_{L^2(D)} \le \ell_D\|\nabla p\|_{\boldsymbol{L}^2(D)}$ for all $p \in H_0^1(D)$. Then (56.1b) follows from $C_{\mathrm{PS}}\ell_D^{-1}\|p\|_{L^2(D)} \le \|\mathrm{d}^{-1}\boldsymbol{\sigma} + \nabla p\|_{\boldsymbol{L}^2(D)} + \|\mathrm{d}^{-1}\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}$, $\|\mathrm{d}^{-1}\boldsymbol{\sigma} + \nabla p\|_{\boldsymbol{L}^2(D)} \le \sup_{w\in L}\frac{|(A(v),w)_L|}{\|w\|_L}$, and $\|\mathrm{d}^{-1}\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)} \le \lambda_\flat^{-1}\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}$, whereas (56.1c) follows from $C_{\mathrm{PS}}\ell_D^{-1}\|q\|_{L^2(D)} \le \lambda_\flat^{-1}\|\boldsymbol{\tau}\|_{\boldsymbol{L}^2(D)}$ with $\mathrm{d}^{-1}\boldsymbol{\tau} + \nabla q = \boldsymbol{0}$.

Exercise 56.6 ((BNB1) for Darcy and Maxwell). (i) For Darcy's equations, we take $(\boldsymbol{\tau}, q) := (\boldsymbol{\sigma} + \mathrm{d}\nabla p, p + \mu^{-1}\nabla\cdot\boldsymbol{\sigma})$. We observe that $\|(\boldsymbol{\tau}, q)\|_L$ is bounded by $\|(\boldsymbol{\sigma}, p)\|_V$ and that

$$(A(\boldsymbol{\sigma}, p), (\boldsymbol{\tau}, q))_L = (\mathrm{d}\boldsymbol{\sigma}, \boldsymbol{\sigma})_{\boldsymbol{L}^2(D)} + (\mathrm{d}\nabla p, \nabla p)_{\boldsymbol{L}^2(D)} + (\mu^{-1}\nabla\cdot\boldsymbol{\sigma}, \nabla\cdot\boldsymbol{\sigma})_{L^2(D)} + (\mu p, p)_{\boldsymbol{L}^2(D)},$$

since $(\boldsymbol{\sigma}, \nabla p)_{\boldsymbol{L}^2(D)} + (\nabla\cdot\boldsymbol{\sigma}, p)_{L^2(D)} = (\boldsymbol{\sigma}\cdot\boldsymbol{n}, p)_{L^2(\partial D)} = 0$ owing to the boundary condition. This allows us to control all the terms composing the graph norm $\|(\boldsymbol{\sigma}, p)\|_V$.

(ii) For Maxwell's equations, we take the test function

$$(\boldsymbol{e}, \boldsymbol{b}) := (e^{-\mathrm{i}\theta}(\boldsymbol{E} - \mathrm{i}\frac{1}{\sigma}\nabla\times\boldsymbol{H}), e^{\mathrm{i}\theta}(\boldsymbol{H} + \frac{1}{\omega\mu}\nabla\times\boldsymbol{E})),$$

where $\theta := \frac{\pi}{4}$. We first observe that $\|(\boldsymbol{e},\boldsymbol{b})\|_L$ is bounded by $\|(\boldsymbol{E},\boldsymbol{H})\|_V$. Recalling that $\tilde\mu := \omega\mu$, $\mathrm{i}e^{-\mathrm{i}\theta} = e^{\mathrm{i}\theta}$, and $\mathrm{i}e^{\mathrm{i}\theta} = -e^{-\mathrm{i}\theta}$, we have

$$\begin{aligned}(A(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{e},\boldsymbol{b}))_L &= e^{\mathrm{i}\theta}(\sigma\|\boldsymbol{E}\|_{\boldsymbol{L}^2(D)}^2 + \tilde\mu\|\boldsymbol{H}\|_{\boldsymbol{L}^2(D)}^2)\\ &\quad + e^{-\mathrm{i}\theta}(\tilde\mu^{-1}\|\nabla\times\boldsymbol{E}\|_{\boldsymbol{L}^2(D)}^2 + \sigma^{-1}\|\nabla\times\boldsymbol{H}\|_{\boldsymbol{L}^2(D)}^2)\\ &\quad + 2\Re\big(e^{\mathrm{i}\theta}\big((\boldsymbol{H},\nabla\times\boldsymbol{E})_{\boldsymbol{L}^2(D)} - (\nabla\times\boldsymbol{H},\boldsymbol{E})_{\boldsymbol{L}^2(D)}\big)\big).\end{aligned}$$

The last term vanishes owing to the boundary condition. This gives

$$\sqrt{2}\Re\left((A(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{e},\boldsymbol{b}))_L\right) = \sigma\|\boldsymbol{E}\|_{\boldsymbol{L}^2(D)}^2 + \tilde\mu\|\boldsymbol{H}\|_{\boldsymbol{L}^2(D)}^2 + \tilde\mu^{-1}\|\nabla\times\boldsymbol{E}\|_{\boldsymbol{L}^2(D)}^2 + \sigma^{-1}\|\nabla\times\boldsymbol{H}\|_{\boldsymbol{L}^2(D)}^2.$$

This allows us to control all the terms composing the graph norm $\|(\boldsymbol{E},\boldsymbol{H})\|_V$.

Exercise 56.7 (Boundary operator for Darcy and Maxwell). (i) (56.7a) holds true since M is skew-symmetric. Moreover, we have

$$\begin{aligned}\langle(M-N)(\boldsymbol{\sigma},p),(\boldsymbol{\tau},q)\rangle_{V',V} &= -2\langle\boldsymbol{\tau}\cdot\boldsymbol{n},p\rangle_{\partial D},\\ \langle(M+N)(\boldsymbol{\sigma},p),(\boldsymbol{\tau},q)\rangle_{V',V} &= 2\langle\boldsymbol{\sigma}\cdot\boldsymbol{n},q\rangle_{\partial D}.\end{aligned}$$

Hence, (56.7b) follows from $(\boldsymbol{\sigma},p) = (\boldsymbol{\sigma},0) + (\boldsymbol{0},p)$, $(\boldsymbol{\sigma},0) \in \ker(M-N)$, and $(\boldsymbol{0},p) \in \ker(M+N)$. Moreover, if $(\boldsymbol{\sigma},p) \in \ker(M-N)$, then $\langle\boldsymbol{\tau}\cdot\boldsymbol{n},p\rangle_{\partial D} = 0$ for all $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div};D)$. Owing to Theorem 4.15, the normal trace operator $\gamma^{\mathrm{d}} : \boldsymbol{H}(\mathrm{div};D) \to H^{-\frac12}(\partial D)$ such that $\gamma^{\mathrm{d}}(\boldsymbol{\tau}) = \boldsymbol{\tau}\cdot\boldsymbol{n}$ is surjective. Since $\boldsymbol{\tau}$ is arbitrary in $\boldsymbol{H}(\mathrm{div};D)$, we conclude that $p_{|\partial D} = 0$. Finally, a Neumann condition can be enforced with the operator

$$\langle M(\boldsymbol{\sigma},p),(\boldsymbol{\tau},q)\rangle_{V',V} := -\langle\boldsymbol{\sigma}\cdot\boldsymbol{n},q\rangle_{\partial D} + \langle\boldsymbol{\tau}\cdot\boldsymbol{n},p\rangle_{\partial D}.$$

(ii) (56.27a) (and (56.7a)) holds true since M is skew-symmetric. Moreover, we have

$$\begin{aligned}\langle(M-N)(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{e},\boldsymbol{b})\rangle_{V',V} &= \int_D \nabla\cdot(2e^{\mathrm{i}\theta}\boldsymbol{H}\times\overline{\boldsymbol{e}})\,\mathrm{d}x,\\ \langle(M+N)(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{e},\boldsymbol{b})\rangle_{V',V} &= \int_D \nabla\cdot(2e^{-\mathrm{i}\theta}\boldsymbol{E}\times\overline{\boldsymbol{b}})\,\mathrm{d}x.\end{aligned}$$

Hence, (56.27b) (and (56.7b)) follows from $(\boldsymbol{E},\boldsymbol{H}) = (\boldsymbol{E},\boldsymbol{0}) + (\boldsymbol{0},\boldsymbol{H})$, $(\boldsymbol{E},\boldsymbol{0}) \in \ker(M-N)$, and $(\boldsymbol{0},\boldsymbol{H}) \in \ker(M+N)$. Moreover, if $(\boldsymbol{E},\boldsymbol{H}) \in \ker(M-N)$, then $2e^{\mathrm{i}\theta}\int_D \nabla\cdot(\boldsymbol{H}\times\overline{\boldsymbol{e}})\,\mathrm{d}x = 0$ for all $\boldsymbol{e} \in \boldsymbol{H}(\mathrm{curl};D)$. Let $\boldsymbol{\phi} \in \boldsymbol{H}^{\frac12}(\partial D)$. Owing to the surjectivity of the trace operator from $\boldsymbol{H}^1(D)$ onto $\boldsymbol{H}^{\frac12}(\partial D)$ applied componentwise, we infer that there is $\boldsymbol{e} \in \boldsymbol{H}^1(D) \subset \boldsymbol{H}(\mathrm{curl};D)$ such that $\boldsymbol{e}_{|\partial D} = \boldsymbol{\phi}$. Using (4.11), we obtain

$$\langle\boldsymbol{H}\times\boldsymbol{n},\boldsymbol{\phi}\rangle_{\partial D} = \int_D \big(\boldsymbol{H}\cdot\nabla\times\overline{\boldsymbol{\phi}} - (\nabla\times\boldsymbol{H})\cdot\overline{\boldsymbol{\phi}}\big)\,\mathrm{d}x = -\int_D \nabla\cdot(\boldsymbol{H}\times\overline{\boldsymbol{\phi}})\,\mathrm{d}x = 0.$$

Since $\boldsymbol{\phi}$ is arbitrary in $\boldsymbol{H}^{\frac12}(\partial D)$, we infer that $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$ in $\boldsymbol{H}^{-\frac12}(\partial D)$. Finally, the condition $\boldsymbol{E}\times\boldsymbol{n} = \boldsymbol{0}$ can be enforced with the operator

$$\langle M(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{e},\boldsymbol{b})\rangle_{V',V} := -\int_D \nabla\cdot(e^{-\mathrm{i}\theta}\boldsymbol{E}\times\overline{\boldsymbol{b}} + e^{\mathrm{i}\theta}\boldsymbol{H}\times\overline{\boldsymbol{e}})\,\mathrm{d}x.$$

Exercise 56.8 (Separation assumption). We have

$$\|u\|_{L^2(D)}^2 = \int_0^1 x_2^{2\alpha}\int_{-x_2}^{x_2}\mathrm{d}x_1\,\mathrm{d}x_2 = 2\int_0^1 x_2^{2\alpha+1}\,\mathrm{d}x_2.$$

Hence, $\|u\|_{L^2(D)} < \infty$ if and only if $\alpha > -1$. Moreover, $\|\boldsymbol{\beta}\cdot\nabla u\|_{L^2(D)} = 0$. Hence, $u \in V$ iff $\alpha > -1$. Finally, observing that $|\boldsymbol{\beta}\cdot\boldsymbol{n}| = \frac{1}{\sqrt{2}}$ and $\mathrm{d}l = \sqrt{2}\,\mathrm{d}x_2$ on $\partial D^{\pm}$, we infer that

$$\|u\|^2_{L^2(|\beta\cdot\boldsymbol{n}|;\partial D)} = \int_{\partial D} u^2|\boldsymbol{\beta}\cdot\boldsymbol{n}|\,\mathrm{d}l = 2\frac{1}{\sqrt{2}}\sqrt{2}\int_0^1 x_2^{2\alpha}\,\mathrm{d}x_2.$$

The integral is finite iff $\alpha > -\frac{1}{2}$.

Exercise 56.9 (Semi-norm $|\cdot|_M$). Notice first that

$$\langle N(v), w\rangle_{V',V} = \langle v, N^*(w)\rangle_{V'',V'} = \overline{\langle N^*(w), v\rangle_{V',V}} = \overline{\langle N(w), v\rangle_{V',V}}.$$

Hence, $\langle N(v), v\rangle_{V',V} \in \mathbb{R}$ for all $v \in V$. Let $t \in \mathbb{R}$, let $v, w \in V_0$. We have

$$\begin{aligned}
0 &\le \Re(\langle M(v+tw), v+tw\rangle_{V',V}) = \langle N(v+tw), v+tw\rangle_{V',V}\\
&= \langle N(v), v\rangle_{V',V} + t\langle N(v), w\rangle_{V',V} + t\langle N(w), v\rangle_{V',V} + t^2\langle N(w), w\rangle_{V',V}\\
&= \langle N(v), v\rangle_{V',V} + 2t\Re\big(\langle N(v), w\rangle_{V',V}\big) + t^2\langle N(w), w\rangle_{V',V}.
\end{aligned}$$

Since this quadratic polynomial in t takes nonnegative values, its discriminant is negative, which implies that

$$\begin{aligned}
\Re\big(\langle N(v), w\rangle_{V',V}\big) &\le \langle N(v), v\rangle^{\frac{1}{2}}_{V',V}\langle N(w), w\rangle^{\frac{1}{2}}_{V',V}\\
&= \Re\big(\langle N(v), v\rangle^{\frac{1}{2}}_{V',V}\big)\Re\big(\langle N(w), w\rangle^{\frac{1}{2}}_{V',V}\big)\\
&= \Re\big(\langle M(v), v\rangle^{\frac{1}{2}}_{V',V}\big)\Re\big(\langle M(w), w\rangle^{\frac{1}{2}}_{V',V}\big) = |v|_M|w|_M.
\end{aligned}$$

There is nothing to prove if $\langle N(v), w\rangle_{V',V} = 0$. Instead, if $\langle N(v), w\rangle_{V',V} \neq 0$, we multiply v by $\overline{\langle N(v), w\rangle_{V',V}}/|\langle N(v), w\rangle_{V',V}|$ in the above formula, and we obtain the expected bound, i.e., $|\langle N(v), w\rangle_{V',V}| \le |v|_M|w|_M$ for all $v, w \in V_0$.

Chapter 57

Residual-based stabilization

Exercises

Exercise 57.1 (Least-squares). Write the LS approximation and the resulting error estimate for the advection-reaction, Darcy's, and Maxwell's equations (for simplicity assume that $u \in H^{k+1}(D;\mathbb{C}^m)$ and hide the scaling factors in the generic constant c).

Exercise 57.2 (Transport in 1D). Consider the LS approximation using $\mathbb{P}_k$ Lagrange finite elements, $k \geq 1$, of the one-dimensional transport problem $u' = f$ in $D := (0,1)$ with $u(0) = 0$ and $f \in H^k(D)$. Prove the optimal L^2-error estimate $\|u - u_h\|_{L^2(D)} \leq ch^{k+1}|f|_{H^k(D)}$. (*Hint*: use a duality argument.)

Exercise 57.3 (Duality argument for Darcy). Consider the LS approximation of Darcy's equations with homogeneous Dirichlet conditions on p in the mixed-order case $k := k_{\boldsymbol{\sigma}} - 1 = k_p \geq 1$, i.e., $V_{h0} := \boldsymbol{P}^{\mathrm{g}}_{k+1}(\mathcal{T}_h) \times P^{\mathrm{g}}_{k,0}(\mathcal{T}_h)$. Assume that $\mu := 0$, $\mathbb{d}^{-1} := \kappa \mathbb{I}_d$ with $\kappa \in W^{1,\infty}(D)$, and that full elliptic regularity holds true for the Laplacian. The goal is to prove the error bound $\|p - p_h\|_{L^2(D)} \leq ch^{k+1}(|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+2}(D)} + |p|_{H^{k+1}(D)})$; see Pehlivanov et al. [38]. Let $\mathcal{I}_h$ have optimal approximation properties in $\boldsymbol{P}^{\mathrm{g}}_{k+1}(\mathcal{T}_h)$, and let $\Pi^{\mathrm{e}}_h : H^1_0(D) \to P^{\mathrm{g}}_{k,0}(\mathcal{T}_h)$ be the elliptic projection such that for all $q \in H^1_0(D)$, $(\nabla(q - \Pi^{\mathrm{e}}_h(q)), \nabla q_h)_{\boldsymbol{L}^2(D)} = 0$ for all $q_h \in P^{\mathrm{g}}_{k,0}(\mathcal{T}_h)$ (see §32.4). (i) Setting $e_h := (\mathcal{I}_h(\boldsymbol{\sigma}) - \boldsymbol{\sigma}_h, \Pi^{\mathrm{e}}_h(p) - p_h)$, prove that $\|e_h\|_V \leq c(\|\mathcal{I}_h(\boldsymbol{\sigma}) - \boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div};D)} + \|\Pi^{\mathrm{e}}_h(p) - p\|_{L^2(D)})$. (*Hint*: use coercivity and the Galerkin orthogonality property.) (ii) Show that $\|p - p_h\|_{L^2(D)} \leq ch^{k+1}(|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+2}(D)} + |p|_{H^{k+1}(D)})$. (*Hint*: use a Poincaré–Steklov inequality and Exercise 32.1.)

Exercise 57.4 (SUPG). Assume that $h_K \leq \beta_K \mu_0^{-1} \min(1, \frac{1}{2}\frac{\mu_0^2}{\mu_\infty^2})$ for all $K \in \mathcal{T}_h$ with $\mu_\infty := \|\mathcal{K}\|_{\mathbb{L}^\infty(D)}$. Prove that the same error estimate as in the GaLS approximation is obtained by considering the following discrete problem: Find $u_h \in V_{h0}$ such that $a_h^{\mathrm{SUPG}}(u_h, w_h) = (f, w_h + \tau A_1(w_h))_L$ for all $w_h \in V_{h0}$ with the SUPG-stabilized sesquilinear form $a_h^{\mathrm{SUPG}}(v_h, w_h) := (A(v_h), w_h)_L + (A(v_h), \tau A_1(w_h))_L$. (*Hint*: bound $(A(v_h), \tau\mathcal{K}v_h)_L$ and use Lemma 57.6 to establish coercivity.)

Exercise 57.5 (Boundary penalty). (i) Prove that (57.33c) and (57.33d) are equivalent. (*Hint*: consider the Hermitian and skew-Hermitian parts of $\mathcal{M}_F$.) (ii) Verify that the boundary penalty operators defined in Example 57.18 for Darcy's equations and in Example 57.19 for Maxwell's equations satisfy (57.33). (*Hint*: direct verification.)

Solution to exercises

Exercise 57.1 (Least-squares). For the advection-reaction equation, the LS approximation amounts to seeking $u_h \in V_{h0}$ such that

$$\int_D (\mu u_h + \boldsymbol{\beta}\cdot\nabla u_h)(\mu w_h + \boldsymbol{\beta}\cdot\nabla w_h)\,\mathrm{d}x = \int_D f(\mu w_h + \boldsymbol{\beta}\cdot\nabla w_h)\,\mathrm{d}x,$$

for all $w_h \in V_{h0}$. Assuming $u \in H^{k+1}(D)$ yields the error estimate

$$\|u - u_h\|_{L^2(D)} + \|\boldsymbol{\beta}\cdot\nabla(u - u_h)\|_{L^2(D)} \le c\,h^k |u|_{H^{k+1}(D)}.$$

For Darcy's equations, the LS approximation amounts to seeking $u_h := (\boldsymbol{\sigma}_h, p_h) \in V_{h0}$ such that

$$\begin{aligned}\int_D d_*(\mathbb{d}^{-1}\boldsymbol{\sigma}_h + \nabla p_h)\cdot(\mathbb{d}^{-1}\boldsymbol{\tau}_h + \nabla q_h)\,\mathrm{d}x \\ + \int_D \mu_*^{-1}(\mu p_h + \nabla\cdot\boldsymbol{\sigma}_h)(\mu q_h + \nabla\cdot\boldsymbol{\tau}_h)\,\mathrm{d}x = \int_D \mu_*^{-1} f(\mu q_h + \nabla\cdot\boldsymbol{\tau}_h)\,\mathrm{d}x,\end{aligned}$$

for all $w_h := (\boldsymbol{\tau}_h, q_h) \in V_{h0}$. Assuming $\boldsymbol{\sigma} \in \boldsymbol{H}^{k+1}(D)$ and $p \in H^{k+1}(D)$ yields

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{H}(\mathrm{div};D)} + \|p - p_h\|_{H^1(D)} \le c\,h^k |(\boldsymbol{\sigma}, p)|_{\boldsymbol{H}^{k+1}(D)\times H^{k+1}(D)}.$$

For Maxwell's equations, the LS approximation amounts to seeking $u_h := (\boldsymbol{E}_h, \boldsymbol{H}_h) \in V_{h0}$ such that

$$\begin{aligned}\int_D \sigma_*^{-1}(\sigma\boldsymbol{E}_h - \nabla\times\boldsymbol{H}_h)\cdot(\sigma\overline{\boldsymbol{e}}_h - \nabla\times\overline{\boldsymbol{b}}_h)\,\mathrm{d}x \\ + \int_D \tilde{\mu}_*^{-1}(\mathrm{i}\omega\mu\boldsymbol{H}_h + \nabla\times\boldsymbol{E}_h)\cdot(-\mathrm{i}\omega\mu\overline{\boldsymbol{b}}_h + \nabla\times\overline{\boldsymbol{e}}_h)\,\mathrm{d}x = \int_D \sigma_*^{-1}\boldsymbol{j}\cdot(\sigma\overline{\boldsymbol{e}}_h - \nabla\times\overline{\boldsymbol{b}}_h)\,\mathrm{d}x,\end{aligned}$$

for all $w_h := (\boldsymbol{e}_h, \boldsymbol{b}_h) \in V_{h0}$. Assuming $(\boldsymbol{E}, \boldsymbol{H}) \in \boldsymbol{H}^{k+1}(D)\times\boldsymbol{H}^{k+1}(D)$ yields

$$\|\boldsymbol{E} - \boldsymbol{E}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} + \|\boldsymbol{H} - \boldsymbol{H}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \le c\,h^k |(\boldsymbol{E}, \boldsymbol{H})|_{\boldsymbol{H}^{k+1}(D)\times\boldsymbol{H}^{k+1}(D)}.$$

Exercise 57.2 (Transport in 1D). The discrete problem amounts to seeking $u_h \in V_{h0}$ such that $\int_D u_h' w_h'\,\mathrm{d}t = \int_D f w_h'\,\mathrm{d}t$ for all $w_h \in V_{h0}$ with

$$V_{h0} := \{v_h \in C^0(\overline{D}) \mid \forall i \in \{0{:}I{-}1\},\, v_{h|[x_i,x_{i+1}]} \in \mathbb{P}_k \mid v_h(0) = 0\}.$$

Consider the adjoint solution $\zeta \in H_0^1(D)$ such that $-\zeta'' = u - u_h$. We have

$$\|u - u_h\|_{L^2(D)}^2 = \int_D (u - u_h)'\zeta'\,\mathrm{d}t = \int_D (u - u_h)'(\zeta - \mathcal{I}_{h0}(\zeta))'\,\mathrm{d}t,$$

where we used the Galerkin orthogonality property and the fact that the Lagrange interpolant of ζ, $\mathcal{I}_{h0}(\zeta)$, is in V_{h0}. Using the Cauchy–Schwarz inequality, the approximation properties of $\mathcal{I}_{h0}$, and the fact that $|\zeta|_{H^2(D)} = \|u - u_h\|_L$, we infer that $\|u - u_h\|_{L^2(D)} \le ch\|(u - u_h)'\|_{L^2(D)}$. Finally, since $f \in H^k(D)$, we have $u \in H^{k+1}(D)$ with $|u|_{H^{k+1}(D)} = |f|_{H^k(D)}$, and (57.7) implies that $\|(u - u_h)'\|_{L^2(D)} \le ch^k|f|_{H^k(D)}$.

Exercise 57.3 (Duality argument for Darcy). (i) Proposition 57.1 implies that

$$\alpha^2\|e_h\|_V^2 \le a^{\mathrm{LS}}(e_h,e_h) = a^{\mathrm{LS}}(\eta_h,e_h),$$

owing to the Galerkin orthogonality property, where $\eta_h \coloneqq (\mathcal{I}_h(\boldsymbol{\sigma})-\boldsymbol{\sigma},\Pi_h^{\mathrm{E}}(p)-p)$. Writing

$$e_h \coloneqq (\boldsymbol{e}_h^\sigma, e_h^p), \qquad \eta_h \coloneqq (\boldsymbol{\eta}_h^\sigma,\eta_h^p),$$

we obtain

$$\begin{aligned}\alpha^2\|e_h\|_V^2 &\le \int_D (\kappa\boldsymbol{\eta}_h^\sigma+\nabla\eta_h^p)\cdot(\kappa\boldsymbol{e}_h^\sigma+\nabla e_h^p)\,\mathrm{d}x + \int_D(\nabla\cdot\boldsymbol{\eta}_h^\sigma)(\nabla\cdot\boldsymbol{e}_h^\sigma)\,\mathrm{d}x\\ &= \int_D(\kappa^2\boldsymbol{\eta}_h^\sigma\cdot\boldsymbol{e}_h^\sigma+\nabla\eta_h^p\cdot\kappa\boldsymbol{e}_h^\sigma+\kappa\boldsymbol{\eta}_h^\sigma\cdot\nabla e_h^p+(\nabla\cdot\boldsymbol{\eta}_h^\sigma)(\nabla\cdot\boldsymbol{e}_h^\sigma))\,\mathrm{d}x\\ &= \int_D(\kappa^2\boldsymbol{\eta}_h^\sigma\cdot\boldsymbol{e}_h^\sigma+\nabla\eta_h^p\cdot\kappa\boldsymbol{e}_h^\sigma-\nabla\cdot(\kappa\boldsymbol{\eta}_h^\sigma)e_h^p+(\nabla\cdot\boldsymbol{\eta}_h^\sigma)(\nabla\cdot\boldsymbol{e}_h^\sigma))\,\mathrm{d}x,\end{aligned}$$

where we used the definition of the elliptic projection and integrated by parts the term involving ∇e_h^p. Using the Cauchy–Schwarz inequality and the smoothness assumption on κ, we obtain

$$\alpha^2\|e_h\|_V^2 \le c(\|\mathcal{I}_h(\boldsymbol{\sigma})-\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div};D)}+\|\Pi_h^{\mathrm{E}}(p)-p\|_{L^2(D)})\|e_h\|_V,$$

and the expected bound follows.
(ii) Using the approximation properties of $\mathcal{I}_h$ and since $k_{\boldsymbol{\sigma}}=k+1$, we infer that

$$\|\mathcal{I}_h(\boldsymbol{\sigma})-\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div};D)} \le ch^{k+1}|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+2}(D)}.$$

Since we are assuming that full elliptic regularity holds true for the Laplacian, Exercise 32.1 shows that $\|\Pi_h^{\mathrm{E}}(p)-p\|_{L^2(D)} \le ch^{k+1}|p|_{H^{k+1}(D)}$. Hence,

$$\|e_h\|_V \le ch^{k+1}(|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+2}(D)}+|p|_{H^{k+1}(D)}).$$

Using the triangle inequality and the Poincaré–Steklov inequality, we infer that

$$\begin{aligned}\|p-p_h\|_{L^2(D)} &\le \|p-\Pi_h^{\mathrm{E}}(p)\|_{L^2(D)}+\|\Pi_h^{\mathrm{E}}(p)-p_h\|_{L^2(D)}\\ &\le \|p-\Pi_h^{\mathrm{E}}(p)\|_{L^2(D)}+C_{\mathrm{PS}}^{-1}\ell_D\|\nabla(\Pi_h^{\mathrm{E}}(p)-p_h)\|_{\boldsymbol{L}^2(D)}\\ &\le \|p-\Pi_h^{\mathrm{E}}(p)\|_{L^2(D)}+C_{\mathrm{PS}}^{-1}\ell_D\|e_h\|_V,\end{aligned}$$

and we conclude using the approximation properties of Π_h^{E} and the above bound on $\|e_h\|_V$.

Exercise 57.4 (SUPG). Using Young's inequality and the assumption on h_K, we infer that for all $v_h\in V_{h0}$,

$$|(A(v_h),\tau\mathcal{K}v_h)_L| \le \frac12\|\tau^{\frac12}A(v_h)\|_L^2+\frac12\|\tau^{\frac12}\mathcal{K}v_h\|_L^2 \le \frac12\|\tau^{\frac12}A(v_h)\|_L^2+\frac12\mu_0\|v_h\|_L^2.$$

Let us denote by a_h^{GL} the discrete sesquilinear form associated with the GaLS approximation and let $\|\cdot\|_{V_{h0}}$ be the stability norm defined in (57.13). This gives

$$\Re\big(a_h^{\mathrm{SUPG}}(v_h,v_h)\big) = \Re\big(a_h^{\mathrm{GL}}(v_h,v_h)\big)-\Re\big((A(v_h),\tau\mathcal{K}v_h)_L\big) \ge \frac12\|v_h\|_{V_{h0}}^2,$$

owing to Lemma 57.6 and the above bound on the nonsymmetric term. Furthermore, the consistency error resulting from a_h^{SUPG} can be estimated by proceeding as in the proof of Lemma 57.7. We conclude that the same error estimate is obtained. As a conclusion, GaLS is more stable than SUPG, and the price that SUPG has to pay for artificially breaking the symmetry of the stabilized sesquilinear form is to require that the meshsize is small enough.

Exercise 57.5 (Boundary penalty). (i) Let us prove that (57.33c) implies (57.33d) (the proof for the converse is similar). Let $\mathcal{M}_F^{\pm} := \frac{1}{2}(\mathcal{M}_F^{\mathsf{H}} \pm \mathcal{M}_F)$ be the Hermitian and skew-Hermitian parts of $\mathcal{M}_F$. We observe that

$$\begin{aligned}
|((\mathcal{M}_F^{\mathrm{BP}} + \mathcal{N}_F)y, z)_{L(F)}| &= |((\mathcal{M}_F + \mathcal{S}_F^{\partial} + \mathcal{N}_F)y, z)_{L(F)}| \\
&\le |((\mathcal{M}_F^{+} + \mathcal{S}_F^{\partial})y, z)_{L(F)}| + |((\mathcal{M}_F^{-} + \mathcal{N}_F)y, z)_{L(F)}| \\
&\le |y|_{\mathcal{M}_F^{\mathrm{BP}}}|z|_{\mathcal{M}_F^{\mathrm{BP}}} + |(y, (\mathcal{M}_F^{-} - \mathcal{N}_F)z)_{L(F)}| \\
&\le 2|y|_{\mathcal{M}_F^{\mathrm{BP}}}|z|_{\mathcal{M}_F^{\mathrm{BP}}} + |((\mathcal{M}_F^{\mathrm{BP}} - \mathcal{N}_F)z, y)_{L(F)}| \\
&\le 2|y|_{\mathcal{M}_F^{\mathrm{BP}}}|z|_{\mathcal{M}_F^{\mathrm{BP}}} + c\beta_{K_l}^{\frac{1}{2}}\|y\|_{L(F)}|z|_{\mathcal{M}_F^{\mathrm{BP}}},
\end{aligned}$$

where we used the triangle inequality to pass to the second line, the fact that $(\mathcal{M}_F^{+} + \mathcal{S}_F^{\partial})$ is Hermitian and positive semidefinite and the Hermitian symmetry of $\mathcal{N}_F$ to pass to the third line, we have added and subtracted $(\mathcal{M}_F^{+} + \mathcal{S}_F^{\partial})$ and proceeded similarly to pass to the fourth line, and we used (57.33c) to pass to the fifth line. We conclude by observing that $|y|_{\mathcal{M}_F^{\mathrm{BP}}} \le c\beta_{K_l}^{\frac{1}{2}}\|y\|_{L(F)}$ owing to (57.33b).

(ii) In both cases, (57.33b) is obvious, so that it remains to prove (57.33a) and (57.33c) (since (57.33d) is equivalent to (57.33c)). Consider Example 57.18. Then $v := (\boldsymbol{\sigma}, p) \in \ker(\mathcal{M}_F - \mathcal{N}_F)$ implies that $p = 0$, so that $v \in \ker(\mathcal{S}_F^{\partial})$. Hence, $v \in \ker(\mathcal{M}_F^{\mathrm{BP}} - \mathcal{N}_F)$. Moreover, we have with $w := (\boldsymbol{\tau}, q)$,

$$((\mathcal{M}_F^{\mathrm{BP}} - \mathcal{N}_F)v, w)_{L(F)} = (\boldsymbol{\tau}\cdot\boldsymbol{n}, p)_{L^2(F)} + \alpha(q, p)_{L^2(F)} \le c\,\|w\|_{L(F)}\|p\|_{L^2(F)}$$

and $\|p\|_{L^2(F)} = \alpha^{-1}|p|_{\mathcal{M}_F^{\mathrm{BP}}}$. Consider now Example 57.19. Then $v := (\boldsymbol{E}, \boldsymbol{H}) \in \ker(\mathcal{M}_F - \mathcal{N}_F)$ implies that $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$, so that $v \in \ker(\mathcal{S}_F^{\partial})$. Hence, $v \in \ker(\mathcal{M}_F^{\mathrm{BP}} - \mathcal{N}_F)$. Moreover, we have with $w := (\boldsymbol{h}, \boldsymbol{e})$,

$$\begin{aligned}
((\mathcal{M}_F^{\mathrm{BP}} - \mathcal{N}_F)v, w)_{L(F)} &= (\boldsymbol{e}, \boldsymbol{H}\times\boldsymbol{n})_{\boldsymbol{L}^2(F)} + \alpha(\boldsymbol{h}\times\boldsymbol{n}, \boldsymbol{H}\times\boldsymbol{n})_{\boldsymbol{L}^2(F)} \\
&\le c\,\|w\|_{L(F)}\|\boldsymbol{H}\times\boldsymbol{n}\|_{\boldsymbol{L}^2(F)},
\end{aligned}$$

and $\|\boldsymbol{H}\times\boldsymbol{n}\|_{\boldsymbol{L}^2(F)} = \alpha^{-1}|\boldsymbol{H}|_{\mathcal{M}_F^{\mathrm{BP}}}$.

Chapter 58

Fluctuation-based stabilization (I)

Exercises

Exercise 58.1 (Simplified setting). Consider the setting of Remark 58.1 and assume that (58.7) holds true. Let $\mathcal{J}_h(v_h) := \frac{h}{\beta}\mathcal{A}_h(v_h)$ for all $v_h \in V_h$. (i) Prove (58.4b). (ii) Prove (58.4c).

Exercise 58.2 (Local bounds for CIP). The goal of this exercise is to prove Lemma 58.4. (i) Let $c_1 \le c_1'$ be positive real numbers. Let a_1, a_2 be two positive real numbers such that $c_1 a_1 \le a_2 \le c_1' a_1$. Verify that there are positive constants c_2, c_2', only depending on c_1 and c_1', such that $c_2 \min(a_1, b) \le \min(a_2, b) \le c_2' \min(a_1, b)$ for any positive real number b. (*Hint*: distinguish the four possible cases.) (ii) Assume (58.19). Prove that there is c such that $\tau_K \le c \min_{K' \in \check{\mathcal{T}}_K^{(2)}} \tau_{K'}$ for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$. (*Hint*: use Step (i) and the regularity of the mesh sequence.) (iii) Prove (58.20). (*Hint*: use Step (ii), $\|\phi\|_{L^\infty(D_K)} \le \max_{L \in \check{\mathcal{T}}_K^{(2)}} \tau_L$, and $\|\phi^{-1}\|_{L^\infty(K)} \le \max_{K' \in \check{\mathcal{T}}_K} \tau_{K'}^{-1}$.)

Exercise 58.3 (Full gradient). Prove (58.21) for CIP with (58.25).

Exercise 58.4 (1D advection, CIP). Let $D := (0,1)$, $f \in L^\infty(D)$, and a nonuniform mesh $\mathcal{T}_h$ of D with nodes $\{x_i\}_{i \in \{0:I+1\}}$ and local cells $K_{i+\frac{1}{2}} := [x_i, x_{i+1}]$ of size $h_{i+\frac{1}{2}} := x_{i+1} - x_i$, $\forall i \in \{0:I\}$. Let $h_i := \frac{1}{2}(h_{i-\frac{1}{2}} + h_{i+\frac{1}{2}})$, $\forall i \in \{1:I\}$, be the length scale associated with the interfaces. Let $V_h := \{v_h \in P_1^{\mathrm{g}}(\mathcal{T}_h) \mid v_h(0) = 0\}$. Let $\beta \neq 0$. Consider the problem $\beta \partial_x u = f$, $u(0) = 0$. (i) Write the CIP formulation for the problem using (58.25) and let $u_h \in V_h$ be the discrete solution. (ii) Show that the discrete problem has a unique solution. (iii) Let $u_h := \sum_{i \in \{1:I+1\}} \mathsf{U}_i \varphi_i$ and $\mathsf{U}_0 := 0$. Write the equation satisfied by $\mathsf{U}_{i-2}, \dots, \mathsf{U}_{i+2}$, $\forall i \in \{2:I-1\}$. (iv) Simplify the equation by assuming that the mesh is uniform and interpret the result in terms of finite differences. (*Hint*: compare the CIP stabilization with the second-order finite difference approximation of $|\beta| h^3 \partial_{xxxx} u$.) *Note*: the term $|\beta| h^3 \partial_{xxxx} u$ is often called *hyperviscosity* in the literature.

Solution to exercises

Exercise 58.1 (Simplified setting). (i) Using the definitions and since $h \le \ell_D$, we have

$$\begin{aligned}\tau^{-\frac12}\|\mathcal{J}_h(v_h)\|_L &= \tau^{\frac12}\|\mathcal{A}_h(v_h)\|_L\\ &\le \beta^{-\frac12}h^{\frac12}\|A_1(v_h)-\mathcal{A}_h(v_h)\|_L + \tau^{\frac12}\|A_1(v_h)\|_L\\ &\le c\,\beta^{-\frac12}h^{\frac12}(\beta^{\frac12}h^{-\frac12}|v_h|_{\mathcal{S}} + \beta\ell_D^{-1}\|v_h\|_L) + \tau^{\frac12}\|A_1(v_h)\|_L\\ &\le \tau^{\frac12}\|A_1(v_h)\|_L + c\,(|v_h|_{\mathcal{S}} + \mu_0^{\frac12}\|v_h\|_L).\end{aligned}$$

This proves (58.4b).
(ii) We have

$$\begin{aligned}2\tau\|A_1(v_h)\|_L^2 &= \tau(A_1(v_h), A_1(v_h)-\tau^{-1}\mathcal{J}_h(v_h))_L + \tau(A_1(v_h)-\tau^{-1}\mathcal{J}_h(v_h), A_1(v_h))_L\\ &\quad + \tau(A_1(v_h), \tau^{-1}\mathcal{J}_h(v_h))_L + \tau(\tau^{-1}\mathcal{J}_h(v_h), A_1(v_h))_L\\ &\le 2\tau\|A_1(v_h)\|_L\|A_1(v_h)-\tau^{-1}\mathcal{J}_h(v_h)\|_L + 2\Re\big((A_1(v_h), \mathcal{J}_h(v_h))_L\big).\end{aligned}$$

Invoking Young's inequality and since $\tau^{-1}\mathcal{J}_h(v_h) := \mathcal{A}_h(v_h)$, we infer that

$$\begin{aligned}\frac12\tau\|A_1(v_h)\|_L^2 &= \Re(A_1(v_h), \mathcal{J}_h(v_h))_L + \frac12\tau\|A_1(v_h)-\mathcal{A}_h(v_h)\|_L^2\\ &\le \Re(A_1(v_h), \mathcal{J}_h(v_h))_L + c\,\tau(\beta h^{-1}|v_h|_{\mathcal{S}}^2 + \beta^2\ell_D^{-2}\|v_h\|_L^2)\\ &\le \Re(A_1(v_h), \mathcal{J}_h(v_h))_L + c\,(|v_h|_{\mathcal{S}}^2 + \mu_0\|v_h\|_L^2),\end{aligned}$$

since $\mu_0 \ge \beta\ell_D^{-1}$. This shows that

$$\frac{1}{2\max(1,c)}\tau\|A_1(v_h)\|_L^2 \le \Re(A_1(v_h), \mathcal{J}_h(v_h))_L + \mu_0\|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2,$$

i.e., (58.4c) holds true.

Exercise 58.2 (Minima). (i) We distinguish four cases.

1) $\min(a_1,b) = \min(a_2,b)$. Then the expected bound holds true with $c_2 := 1$ and $c_2' := 1$.

2) $\min(a_1,b) = b$ and $\min(a_2,b) = a_2$. Then the expected bound holds true with $c_2 := c_1$ and $c_2' := 1$.

3) $\min(a_1,b) = a_1$ and $\min(a_2,b) = b$. Then the expected bound holds true with $c_2 := 1$ and $c_2' := c_1'$.

4) $\min(a_1,b) = a_1$ and $\min(a_2,b) = a_2$. Then the expected bound holds true with $c_2 := c_1$ and $c_2' := c_1'$.

(ii) Let $K \in \mathcal{T}_h$ and let K' be arbitrary in $\check{\mathcal{T}}_K^{(2)}$. The regularity of the mesh sequence implies that $h_K \le ch_{K'}$. Moreover, the assumption (58.19) on the grading of the coefficients $\{\beta_K\}_{K\in\mathcal{T}_h}$ implies that $\beta_K^{-1} \le c\beta_{K'}^{-1}$. Hence, $\beta_K^{-1}h_K \le c\beta_{K'}^{-1}h_{K'}$. Owing to Step (i) we infer that $\tau_K \le c'\tau_{K'}$. Taking the infimum over $K' \in \check{\mathcal{T}}_K^{(2)}$ proves the assertion.
(iii) Let us prove (58.20a). Let $K \in \mathcal{T}_h$. The definition of ϕ implies that

$$\|\phi\|_{L^\infty(D_K)} \le \max_{L\in\check{\mathcal{T}}_K^{(2)}} \tau_L.$$

Let $L \in \check{\mathcal{T}}_K^{(2)}$. Then $\tau_L \le c\tau_K$ owing to Step (ii) since $K \in \check{\mathcal{T}}_L^{(2)}$. Moreover, still owing to Step (ii), we have $\tau_K \le c \min_{K' \in \check{\mathcal{T}}_K} \tau_{K'}$. Combining the above bounds shows that $\|\phi\|_{L^\infty(D_K)} \le c \min_{K' \in \check{\mathcal{T}}_K} \tau_{K'}$. This proves (58.20a).
Let us now prove (58.20b). The definition of ϕ implies that

$$\|\phi^{-1}\|_{L^\infty(K)} \le \max_{K' \in \check{\mathcal{T}}_K} \tau_{K'}^{-1}.$$

Moreover, owing to Step (ii) we have

$$\max_{K' \in \check{\mathcal{T}}_K} \tau_{K'}^{-1} \le \max_{K' \in \check{\mathcal{T}}_K^{(2)}} \tau_{K'}^{-1} \le c\,\tau_K^{-1}.$$

This shows that (58.20b) holds true.

Exercise 58.3 (Full gradient). Owing to the triangle inequality, we infer that

$$\begin{aligned}\sum_{F \in \mathcal{F}_h^\circ} \tau_F h_F \|[\![\underline{A}_1(v_h)]\!]_F\|_{L(F)}^2 &\le 2 \sum_{F \in \mathcal{F}_h^\circ} \tau_F h_F \|[\![A_1(v_h)]\!]_F\|_{L(F)}^2 + \tau_F h_F \|[\![(A_1 - \underline{A}_1)(v_h)]\!]_F\|_{L(F)}^2 \\ &\le 2 \sum_{F \in \mathcal{F}_h^\circ} \tau_F h_F \|[\![A_1(v_h)]\!]_F\|_{L(F)}^2 + c\,\mu_0 \|v_h\|_L^2,\end{aligned}$$

where we bounded the second term on the right-hand side by invoking the triangle inequality to bound the jump, the fact that the fields $\{\mathcal{A}^k\}_{k \in \{1:d\}}$ are piecewise Lipschitz with $L_{\mathcal{A}} \le c\mu_0$, a discrete trace inequality, an inverse inequality, and the fact that $\tau_F \le c \min(\tau_{K_l}, \tau_{K_r})$ and $\mu_0 \le \tau_K^{-1}$. Concerning the first term on the right-hand side, we use the assumption that the fields $\{\mathcal{A}^k\}_{k \in \{1:d\}}$ are continuous over $\overline{D}$ to infer that

$$\sum_{F \in \mathcal{F}_h^\circ} \tau_F h_F \|[\![A_1(v_h)]\!]_F\|_{L(F)}^2 \le \sum_{F \in \mathcal{F}_h^\circ} \tau_F \beta_F^2 h_F \|[\![\nabla v_h]\!]_F\|_{\boldsymbol{L}(F)}^2,$$

with $\beta_F := \max(\beta_{K_l}, \beta_{K_r})$. Finally, we have

$$\begin{aligned}\beta_F \tau_F &\le \max(\beta_{K_l}, \beta_{K_r}) \max(\beta_{K_l}^{-1} h_{K_l}, \beta_{K_r}^{-1} h_{K_r}) \\ &\le c\, h_F \max(\beta_{K_l}, \beta_{K_r}) \min(\beta_{K_l}, \beta_{K_r})^{-1} \le c'\, h_F,\end{aligned}$$

where we used the regularity of the mesh sequence and the assumption on the variations of β which implies that $\max(\beta_{K_l}, \beta_{K_r}) \le c \min(\beta_{K_l}, \beta_{K_r})$. We conclude that

$$\sum_{F \in \mathcal{F}_h^\circ} \tau_F h_F \|[\![A_1(v_h)]\!]_F\|_{L(F)}^2 \le c \sum_{F \in \mathcal{F}_h^\circ} \beta_F h_F^2 \|[\![\nabla v_h]\!]_F\|_{\boldsymbol{L}(F)}^2.$$

This proves the assertion.

Exercise 58.4 (1D advection, CIP). (i) The CIP formulation consists of seeking $u_h \in V_h$ s.t.

$$\int_D v_h \beta \partial_x u_h \,\mathrm{d}x + \sum_{i \in \{1:I\}} |\beta| h_i^2 [\![\partial_x u_h]\!][\![\partial_x v_h]\!] = \int_D f v_h \,\mathrm{d}x, \qquad \forall v_h \in V_h.$$

(ii) We establish uniqueness. Assume $f = 0$. Then, using $v_h := u_h$, we obtain

$$0 = \int_D \frac{1}{2} \beta \partial_x u_h^2 \,\mathrm{d}x + \sum_{i \in \{1:I\}} |\boldsymbol{\beta}| h_i^2 [\![\partial_x u_h]\!]^2 = \frac{1}{2} \beta u_h(1)^2 + \sum_{i \in \{1:I\}} |\beta| h_i^2 [\![\partial_x u_h]\!]^2.$$

This implies that $[\![\partial_x u_h]\!] = 0$ for all $i \in \{1{:}I\}$. Hence, $\partial_x u_h$ is constant over D since u_h is piecewise linear. We conclude that $\partial_x u_h = 0$ because $\partial_x u_h \int_D v_h \,\mathrm{d}x = 0$ for all $v_h \in V_h$. Hence, $u_h(x) = u_h(0) = 0$ for all $x \in D$, thereby proving uniqueness. Existence follows from the fact that the trial and test spaces have the same dimension (they are actually identical).
(iii) Let us now use the shape function φ_i as a test function, for all $i \in \{2{:}I{-}1\}$. We obtain

$$\begin{aligned}\int_D f\varphi_i \,\mathrm{d}x = {}& \int_{x_{i-1}}^{x_i} \varphi_i \beta \frac{\mathsf{U}_i - \mathsf{U}_{i-1}}{h_{i-\frac12}} \,\mathrm{d}x + \int_{x_i}^{x_{i+1}} \varphi_i \beta \frac{\mathsf{U}_{i+1} - \mathsf{U}_i}{h_{i+\frac12}} \,\mathrm{d}x \\ & + |\beta| h_{i-1}^2 \Big(\frac{\mathsf{U}_i - \mathsf{U}_{i-1}}{h_{i-\frac12}} - \frac{\mathsf{U}_{i-1} - \mathsf{U}_{i-2}}{h_{i-\frac32}} \Big) \frac{1}{h_{i-\frac12}} \\ & + |\beta| h_i^2 \Big(\frac{\mathsf{U}_{i+1} - \mathsf{U}_i}{h_{i+\frac12}} - \frac{\mathsf{U}_i - \mathsf{U}_{i-1}}{h_{i-\frac12}} \Big) \Big(- \frac{1}{h_{i+\frac12}} - \frac{1}{h_{i-\frac12}} \Big) \\ & + |\beta| h_{i+1}^2 \Big(\frac{\mathsf{U}_{i+2} - \mathsf{U}_{i+1}}{h_{i+\frac32}} - \frac{\mathsf{U}_{i+1} - \mathsf{U}_i}{h_{i+\frac12}} \Big) \frac{1}{h_{i+\frac12}}.\end{aligned}$$

Since $\int_{x_{i-1}}^{x_i} \varphi_i \,\mathrm{d}x = h_{i-\frac12}$ and $\int_{x_i}^{x_{i+1}} \varphi_i \,\mathrm{d}x = h_{i+\frac12}$, this gives

$$\begin{aligned}\int_D f\varphi_i \,\mathrm{d}x = {}& \beta \frac{\mathsf{U}_{i+1} - \mathsf{U}_{i-1}}{2} \\ & + |\beta| h_{i-1}^2 \Big(\frac{\mathsf{U}_i - \mathsf{U}_{i-1}}{h_{i-\frac12}} - \frac{\mathsf{U}_{i-1} - \mathsf{U}_{i-2}}{h_{i-\frac32}} \Big) \frac{1}{h_{i-\frac12}} \\ & + |\beta| h_i^2 \Big(\frac{\mathsf{U}_{i+1} - \mathsf{U}_i}{h_{i+\frac12}} - \frac{\mathsf{U}_i - \mathsf{U}_{i-1}}{h_{i-\frac12}} \Big) \Big(- \frac{1}{h_{i+\frac12}} - \frac{1}{h_{i-\frac12}} \Big) \\ & + |\beta| h_{i+1}^2 \Big(\frac{\mathsf{U}_{i+2} - \mathsf{U}_{i+1}}{h_{i+\frac32}} - \frac{\mathsf{U}_{i+1} - \mathsf{U}_i}{h_{i+\frac12}} \Big) \frac{1}{h_{i+\frac12}}.\end{aligned}$$

(iv) Assuming that the mesh is uniform, we obtain

$$\begin{aligned}\frac{1}{h} \int_D f\varphi_i \,\mathrm{d}x = {}& \beta \frac{\mathsf{U}_{i+1} - \mathsf{U}_{i-1}}{2h} + \frac{|\beta|}{h} \Big((\mathsf{U}_i - 2\mathsf{U}_{i-1} + \mathsf{U}_{i-2}) \\ & - 2(\mathsf{U}_{i+1} - 2\mathsf{U}_i + \mathsf{U}_{i-1}) + (\mathsf{U}_{i+2} - 2\mathsf{U}_{i+1} + \mathsf{U}_i) \Big).\end{aligned}$$

The term $\beta \frac{\mathsf{U}_{i+1} - \mathsf{U}_{i-1}}{2}$ is the second-order finite difference approximation of $\beta \partial_x u$. The term $\frac{|\beta|}{h}((\mathsf{U}_i - 2\mathsf{U}_{i-1} + \mathsf{U}_{i-2}) - 2(\mathsf{U}_{i+1} - 2\mathsf{U}_i + \mathsf{U}_{i-1}) + (\mathsf{U}_{i+2} - 2\mathsf{U}_{i+1} + \mathsf{U}_i))$ is the second-order finite difference approximation of $|\beta| h^3 \partial_{xxxx} u$. This shows that the CIP formulation amounts to approximating the solution to the perturbed equation $\beta \partial_x u + |\beta| h^3 \partial_{xxxx} u = f$.

Chapter 59

Fluctuation-based stabilization (II)

Exercises

Exercise 59.1 (Inf-sup condition). Consider the setting of §59.1 and assume that the functions in B_h vanish on ∂D. Prove that there is $\alpha > 0$ such that for all $r_h \in R_h$ and all $h \in \mathcal{H}$,

$$\alpha(\|r_h\|_{V_h} + \mu_0^{-\frac{1}{2}}\|A_1(r_h)\|_L) \le \sup_{w_h \in V_h} \frac{|a_h^{\mathrm{BP}}(r_h, w_h)|}{\|w_h\|_{V_h}},$$

with a_h^{BP} defined in (58.3) and $\|v_h\|_{V_h}^2 := \mu_0\|v_h\|_L^2 + \frac{1}{2}|v_h|_{\mathcal{M}}^2 + |v_h|_{\mathcal{S}^\partial}^2$ for all $v_h \in V_h$. (*Hint*: use the coercivity of a_h^{BP} to control $\|r_h\|_{V_h}$, and use that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are piecewise Lipschitz together with (59.4) to control $\mu_0^{-\frac{1}{2}}\|A_1(r_h)\|_L$.)

Exercise 59.2 (Full gradient). Prove (59.9) for the choice of s_h^{LPS} in Example 59.4.

Exercise 59.3 (SGV). Prove Lemma 59.7.

Solution to exercises

Exercise 59.1 (Inf-sup condition). Let us set

$$\rho_h := \sup_{w_h \in V_h} \frac{|a(r_h, w_h)|}{\|w_h\|_{V_h}}.$$

Let $r_h \in R_h$. The coercivity property of a_h^{BP} and the fact that $R_h \subset V_h$ imply that $\|r_h\|_{V_h} \le \rho_h$. Considering the first-order operator $\underline{A}_1$ defined in Proposition 58.5, recalling that the fields $\mathcal{A}^k$ are piecewise Lipschitz, and using an inverse inequality, we obtain that

$$\|(A_1 - \underline{A}_1)(r_h)\|_{L(K)} \le c\,\mu_0\|r_h\|_{L(K)},$$

for all $K \in \mathcal{T}_h$. Since $(\underline{A}_1(r_h))_{|K} \in G_K$, the assumption (59.4) implies that

$$\begin{aligned}\|\underline{A}_1(r_h)\|_{L(K)} &\le \gamma^{-1} \sup_{b_K \in B_K} \frac{|(\underline{A}_1(r_h), b_K)_{L(K)}|}{\|b_K\|_{L(K)}} \\ &\le \gamma^{-1} \sup_{b_K \in B_K} \frac{|(A_1(r_h), b_K)_{L(K)}|}{\|b_K\|_{L(K)}} + c\,\mu_0 \|r_h\|_{L(K)}.\end{aligned}$$

Since $B_h = \bigoplus_{K\in\mathcal{T}_h} B_K$, taking a supremum over functions in B_h is achieved by taking suprema over functions in B_K independently for all $K \in \mathcal{T}_h$. This implies that there is $c > 0$ s.t. for all $h \in \mathcal{H}$,

$$c\,\mu_0^{-\frac{1}{2}} \|\underline{A}_1(r_h)\|_L \le \sup_{b_h \in B_h} \frac{|(A_1(r_h), b_h)_L|}{\mu_0^{\frac{1}{2}} \|b_h\|_L} + \mu_0^{\frac{1}{2}} \|r_h\|_L.$$

Since $(A_1(r_h), b_h)_L = a_h^{\mathrm{BP}}(r_h, b_h) - (\mathcal{K} r_h, b_h)_L$ owing to the assumption that functions in B_h vanish on ∂D, we infer that

$$c\,\mu_0^{-\frac{1}{2}} \|\underline{A}_1(r_h)\|_L \le \sup_{b_h \in B_h} \frac{|a_h^{\mathrm{BP}}(r_h, b_h)|}{\mu_0^{\frac{1}{2}} \|b_h\|_L} + \mu_0^{\frac{1}{2}} \|r_h\|_L.$$

We conclude by observing that $\mu_0^{\frac{1}{2}} \|b_h\|_L = \|b_h\|_{V_h}$ and $B_h \subset V_h$.

Exercise 59.2 (Full gradient). We observe that

$$\begin{aligned}\|\tau^{\frac{1}{2}} \kappa_h^{\mathrm{G}}(\underline{A}_1(v_h))\|_{L(K)} &\le \sum_{k\in\{1:d\}} \tau_K^{\frac{1}{2}} \|\underline{\mathcal{A}}_K^k \kappa_h^{\mathrm{G}}(\partial_k v_h)\|_{L(K)} \\ &\le c\,\tau_K^{\frac{1}{2}} \beta_K \|\kappa_h^{\mathrm{G}}(\nabla v_h)\|_{\boldsymbol{L}(K)},\end{aligned}$$

for all $K \in \mathcal{T}_h$, where we used the linearity of κ_h^{G}, the triangle inequality, and the bound $\|\underline{\mathcal{A}}_K^k\|_{\ell^2} \le \beta_K$.

Exercise 59.3 (SGV). We have to prove that the bilinear form $s_h^{\mathrm{SGV}}(v_h, w_h)$ defined in (59.19) satisfies the design conditions (58.4).
(1) Let us first prove that (58.4a) holds true. Using an inverse inequality and the inequality $\beta_K \tau_K \le h_K$, we obtain

$$|s_h^{\mathrm{SGV}}(v_h, v_h)| \le c \sum_{K\in\mathcal{T}_h} \tau_K \|\kappa_h^{\mathrm{R}}(v_h)\|_{L(K)}^2.$$

Now, we use the local stability of κ_h^{R} (which follows from $\|\pi_h^{\mathrm{R}}(v_h)\|_{L(K)} \le c\|v_h\|_{L(\tilde{D}_K)}$, see (59.16)) and the regularity of the mesh sequence to infer that

$$|s_h^{\mathrm{SGV}}(v_h, v_h)| \le c \sum_{K\in\mathcal{T}_h} \tau_K^{-1} \|v_h\|_{L(\tilde{D}_K)}^2 \le c'\,\|\tau^{-\frac{1}{2}} v_h\|_L^2.$$

(2) We now prove that (59.17) holds true and then invoke Proposition 59.6 to establish (58.4b)-(58.4c). The triangle inequality implies that

$$\|\tau^{\frac{1}{2}} \kappa_h^{\mathrm{R}}(A_1(v_h))\|_L \le \|\tau^{\frac{1}{2}} \kappa_h^{\mathrm{R}}(\underline{A}_1(v_h))\|_L + \|\tau^{\frac{1}{2}} \kappa_h^{\mathrm{R}}((A_1 - \underline{A}_1)(v_h))\|_L.$$

Using the local L-stability of κ_h^{R}, the Lipschitz continuity of the fields $\mathcal{A}^k$, the fact that $L_{\mathcal{A}} \le c\mu_0$ and $\tau_K \le \mu_0^{-1}$, an inverse inequality, and the regularity of the mesh sequence, we obtain (as usual, the value of c changes at each occurrence)

$$\begin{aligned}\|\tau^{\frac{1}{2}}\kappa_h^{\mathrm{R}}((A_1 - \underline{A}_1)(v_h))\|_L^2 &\le c \sum_{K\in\mathcal{T}_h} \tau_K \|(A_1 - \underline{A}_1)(v_h)\|_{L(\tilde{D}_K)}^2 \\ &\le c \sum_{K\in\mathcal{T}_h} \tau_K L_{\mathcal{A}}^2 h_K^2 \|\nabla v_h\|_{\boldsymbol{L}(K)}^2 \\ &\le c \sum_{K\in\mathcal{T}_h} \tau_K L_{\mathcal{A}}^2 \|v_h\|_{L(K)}^2 \le c\,\mu_0 \|v_h\|_L^2.\end{aligned}$$

Hence, $\|\tau^{\frac{1}{2}}\kappa_h^{\mathrm{R}}(A_1(v_h))\|_L \le \|\tau^{\frac{1}{2}}\kappa_h^{\mathrm{R}}(\underline{A}_1(v_h))\|_L + c\mu_0^{\frac{1}{2}}\|v_h\|_L$ for all $v_h \in V_h$. We have proved that (59.17) holds true, and Proposition 59.6 implies that (58.4b) and (58.4c) also hold true.

Chapter 60

Discontinuous Galerkin

Exercises

Exercise 60.1 (Upwind flux). Consider the advection equation $\mu u + \boldsymbol{\beta}\cdot\nabla u = f$. Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$. Let $\widehat{\Phi}_F^{\mathrm{stb}}(u_h) := \boldsymbol{\beta}\cdot\boldsymbol{n}_F\{u_h\} + \frac{1}{2}|\boldsymbol{\beta}\cdot\boldsymbol{n}_F|[\![u_h]\!]$. Show that $\widehat{\Phi}_F^{\mathrm{stb}}(u_h) = (\boldsymbol{\beta}\cdot\boldsymbol{n}_F)u_{h|K_l}$ if $\boldsymbol{\beta}\cdot\boldsymbol{n}_F \geq 0$ and $\widehat{\Phi}_F^{\mathrm{stb}}(u_h) = (\boldsymbol{\beta}\cdot\boldsymbol{n}_F)u_{h|K_r}$ otherwise.

Exercise 60.2 ($\mathcal{S}_F^\circ$). Verify that the jump penalty operators from §60.3.3 verify (60.21).

Exercise 60.3 (Absolute value). (i) Show that a suitable choice for the jump penalty operator is $\mathcal{S}_F^\circ = |\mathcal{N}_F|$ where $|\mathcal{N}_F|$ is the unique Hermitian positive semidefinite matrix such that $|\mathcal{N}_F|^2 = \mathcal{N}_F^{\mathsf{H}}\mathcal{N}_F = \mathcal{N}_F^2$. (*Hint*: $\left|w^{\mathsf{H}}\mathcal{N}_F v\right| \leq \left|w^{\mathsf{H}}|\mathcal{N}_F|v\right|$.) (ii) Verify that

$$\left|\left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \boldsymbol{n}_F \\ \hdashline \boldsymbol{n}_F^{\mathsf{T}} & 0 \end{array}\right]\right| = \left[\begin{array}{c:c} \boldsymbol{n}_F\otimes\boldsymbol{n}_F & \mathbf{0} \\ \hdashline \mathbf{0}^{\mathsf{T}} & 1 \end{array}\right], \quad \left|\left[\begin{array}{c:c} \alpha\mathbb{T}^{\mathsf{T}}\mathbb{T} & \mathbb{O}_{3\times 3} \\ \hdashline \mathbb{O}_{3\times 3} & \beta\mathbb{T}^{\mathsf{T}}\mathbb{T} \end{array}\right]\right| = \left[\begin{array}{c:c} |\alpha|\mathbb{T}^{\mathsf{T}}\mathbb{T} & \mathbb{O}_{3\times 3} \\ \hdashline \mathbb{O}_{3\times 3} & |\beta|\mathbb{T}^{\mathsf{T}}\mathbb{T} \end{array}\right].$$

Exercise 60.4 (Matrix $\mathbb{T}$). (i) Show that $\mathbb{T}^{\mathsf{T}} = -\mathbb{T}$. (ii) Show that $(\mathbb{T}^{\mathsf{T}}\mathbb{T})^2 = \mathbb{T}^{\mathsf{T}}\mathbb{T}$.

Exercise 60.5 (Orthogonal subscales). (i) Prove that a_h^{stb} is coercive on V_h equipped with the norm $\|v_h\|_{V_h}^2 := \mu_0\|v_h\|_L^2 + \frac{1}{2}|v_h|_{\mathcal{M}^{\mathrm{BP}}}^2 + |[\![v_h]\!]|_{\mathcal{S}^\circ}^2$. (ii) Assume that the fields $\mathcal{A}^k$ are Lipschitz (with Lipschitz constant $L_{\mathcal{A}} \leq c\mu_0$). Assume that $u \in V_{\mathrm{S}} := H^s(D;\mathbb{C}^m) \cap V$, $s > \frac{1}{2}$. Prove that there is c such that

$$|\langle \delta_h(\mathcal{I}_h^{\flat}(u)), w_h\rangle_{V_h',V_h}| \leq c\,\|u - \mathcal{I}_h^{\flat}(u)\|_{V_\sharp}\|w_h\|_{V_h},$$

for all $(v, w_h) \in V_\sharp \times V_h$ and all $h \in \mathcal{H}$, where $\mathcal{I}_h^{\flat}$ denotes the L-orthogonal projection onto V_h, $\|v\|_{V_\flat}^2 := \mu_0\|v\|_L^2 + \frac{1}{2}|v|_{\mathcal{M}^{\mathrm{BP}}}^2 + |[\![v]\!]|_{\mathcal{S}^\circ}^2$, and $\|v\|_{V_\sharp}^2 := \|v\|_{V_\flat}^2 + \sum_{K\in\mathcal{T}_h}\beta_K\|v\|_{L(\partial K)}^2$. (*Hint*: adapt the proof of Lemma 60.10.) (iii) Prove that $\|u - u_h\|_{V_\flat} \leq c\phi^{\frac{1}{2}}h^{k+\frac{1}{2}}|u|_{H^{k+1}(D;\mathbb{C}^m)}$ using only Steps (i) and (ii). (*Hint*: adapt the proof of Lemma 27.8.)

Solution to exercises

Exercise 60.1 (Upwind flux). By definition, we have

$$\widehat{\Phi}_F^{\mathrm{stb}}(u_h) = \frac{1}{2}(\boldsymbol{\beta}\cdot\boldsymbol{n}_F)(u_{h|K_l} + u_{h|K_r}) + \frac{1}{2}|\boldsymbol{\beta}\cdot\boldsymbol{n}_F|(u_{h|K_l} - u_{h|K_r}).$$

Assuming that $\boldsymbol{\beta}\cdot\boldsymbol{n}_F \geq 0$, we infer that

$$\begin{aligned}\widehat{\Phi}_F^{\mathrm{stb}}(u_h) &= \frac{1}{2}(\boldsymbol{\beta}\cdot\boldsymbol{n}_F)(u_{h|K_l} + u_{h|K_r}) + \frac{1}{2}(\boldsymbol{\beta}\cdot\boldsymbol{n}_F)(u_{h|K_l} - u_{h|K_r})\\ &= (\boldsymbol{\beta}\cdot\boldsymbol{n}_F)\frac{1}{2}(u_{h|K_l} + u_{h|K_r} + u_{h|K_l} - u_{h|K_r}) = (\boldsymbol{\beta}\cdot\boldsymbol{n}_F)u_{h|K_l}.\end{aligned}$$

The proof is similar if $\boldsymbol{\beta}\cdot\boldsymbol{n}_F \leq 0$.

Exercise 60.2 ($\mathcal{S}_F^\circ$). In all cases, (60.21b) is obvious.
(1) For the advection-reaction equation, $\ker(\mathcal{N}_F) = \{0\} = \ker(\mathcal{S}_F^\circ)$ unless $\boldsymbol{\beta}\cdot\boldsymbol{n}_F = 0$, in which case $\ker(\mathcal{N}_F) = \mathbb{R} = \ker(\mathcal{S}_F^\circ)$. Hence (60.21a) holds true. Moreover, $|(\mathcal{N}_F v, w)_{L(F)}| \leq \alpha^{-2}|v|_{\mathcal{S}_F^\circ}|w|_{\mathcal{S}_F^\circ}$, so that (60.21c) holds true.
(2) For Darcy's equations, $v := (\boldsymbol{\sigma}, p) \in \ker(\mathcal{N}_F)$ implies that $\boldsymbol{\sigma}\cdot\boldsymbol{n} = 0$ and $p = 0$. Hence, (60.21a) holds true. Moreover, we have with $w := (\boldsymbol{\tau}, q)$,

$$(\mathcal{N}_F v, w)_{L(F)} = (\boldsymbol{\sigma}\cdot\boldsymbol{n}, q)_{L^2(F)} + (p, \boldsymbol{\tau}\cdot\boldsymbol{n})_{L^2(F)}.$$

Since $|v|_{\mathcal{S}_F^\circ} = \alpha_1\|\boldsymbol{\sigma}\cdot\boldsymbol{n}_F\|_{L^2(F)} + \alpha_2\|p\|_{L^2(F)}$, we infer that (60.21c) holds true.
(3) For Maxwell's equations, $v := (\boldsymbol{E}, \boldsymbol{H}) \in \ker(\mathcal{N}_F)$ implies that $\boldsymbol{H}\times\boldsymbol{n}_F = \boldsymbol{E}\times\boldsymbol{n}_F = \boldsymbol{0}$. Hence, (60.21a) holds true. Moreover, we have with $w := (\boldsymbol{e}, bh)$,

$$(\mathcal{N}_F v, w)_{L(F)} = (\boldsymbol{H}\times\boldsymbol{n}_F, \boldsymbol{e})_{\boldsymbol{L}^2(F)} + (\boldsymbol{E}, \boldsymbol{h}\cdot\boldsymbol{n}_F)_{\boldsymbol{L}^2(F)}.$$

Since $|v|_{\mathcal{S}_F^\circ} = \alpha_1\|\boldsymbol{E}\times\boldsymbol{n}_F\|_{\boldsymbol{L}^2(F)} + \alpha_2\|\boldsymbol{H}\times\boldsymbol{n}_F\|_{\boldsymbol{L}^2(F)}$, we infer that (60.21c) holds true.

Exercise 60.3 (Penalty field by absolute value). (i) Let us verify that $\mathcal{S}_F^\circ = |\mathcal{N}_F|$ verifies (60.21). (60.21a) is obvious. To prove (60.21b), we use that $\|\,|\mathcal{N}_F|\,\|_{\ell^2} = \|\mathcal{N}_F\|_{\ell^2}$. To prove (60.21c), we use the hint, the Cauchy–Schwarz inequality, and (60.21b).
(ii) A direct calculation shows that

$$\left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \boldsymbol{n}_F \\ \hdashline \boldsymbol{n}_F^\mathsf{T} & 0 \end{array}\right]\left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \boldsymbol{n}_F \\ \hdashline \boldsymbol{n}_F^\mathsf{T} & 0 \end{array}\right] = \left[\begin{array}{c:c} \boldsymbol{n}_F\otimes\boldsymbol{n}_F & \boldsymbol{0} \\ \hdashline \boldsymbol{0}^\mathsf{T} & 1 \end{array}\right] = \left[\begin{array}{c:c} \boldsymbol{n}_F\otimes\boldsymbol{n}_F & \boldsymbol{0} \\ \hdashline \boldsymbol{0}^\mathsf{T} & 1 \end{array}\right]\left[\begin{array}{c:c} \boldsymbol{n}_F\otimes\boldsymbol{n}_F & \boldsymbol{0} \\ \hdashline \boldsymbol{0}^\mathsf{T} & 1 \end{array}\right].$$

Moreover, since the matrix

$$\left[\begin{array}{c:c} \alpha\mathbb{T}^\mathsf{T}\mathbb{T} & \mathbb{O}_{3\times3} \\ \hdashline \mathbb{O}_{3\times3} & \beta\mathbb{T}^\mathsf{T}\mathbb{T} \end{array}\right]$$

is block diagonal, we have

$$\left|\left[\begin{array}{c:c} \alpha\mathbb{T}^\mathsf{T}\mathbb{T} & \mathbb{O}_{3\times3} \\ \hdashline \mathbb{O}_{3\times3} & \beta\mathbb{T}^\mathsf{T}\mathbb{T} \end{array}\right]\right| = \left[\begin{array}{c:c} |\alpha\mathbb{T}^\mathsf{T}\mathbb{T}| & \mathbb{O}_{3\times3} \\ \hdashline \mathbb{O}_{3\times3} & |\beta\mathbb{T}^\mathsf{T}\mathbb{T}| \end{array}\right] = \left[\begin{array}{c:c} |\alpha||\mathbb{T}^\mathsf{T}\mathbb{T}| & \mathbb{O}_{3\times3} \\ \hdashline \mathbb{O}_{3\times3} & |\beta||\mathbb{T}^\mathsf{T}\mathbb{T}| \end{array}\right].$$

But $\mathbb{T}^\mathsf{T}\mathbb{T}$ is symmetric positive semidefinite. Hence, $|\mathbb{T}^\mathsf{T}\mathbb{T}| = \mathbb{T}^\mathsf{T}\mathbb{T}$. This proves the assertion.

Exercise 60.4 (Matrix $\mathbb{T}$). (i) Let $\boldsymbol{E},\boldsymbol{H}\in\mathbb{R}^3$. We have

$$\boldsymbol{E}^{\mathsf{T}}\mathbb{T}\boldsymbol{H}=\boldsymbol{E}\cdot(\boldsymbol{H}\times\boldsymbol{n})=-(\boldsymbol{E}\times\boldsymbol{n})\cdot\boldsymbol{H}=-(\mathbb{T}\boldsymbol{E})^{\mathsf{T}}\boldsymbol{H}=-\boldsymbol{E}^{\mathsf{T}}\mathbb{T}^{\mathsf{T}}\boldsymbol{H}.$$

This proves that $\mathbb{T}^{\mathsf{T}}=-\mathbb{T}$.
(ii) Let $\boldsymbol{E},\boldsymbol{H}\in\mathbb{R}^3$. Since $\mathbb{T}^{\mathsf{T}}=-\mathbb{T}$, we have

$$\begin{aligned}\boldsymbol{E}^{\mathsf{T}}\mathbb{T}^{\mathsf{T}}\mathbb{T}\mathbb{T}^{\mathsf{T}}\mathbb{T}\boldsymbol{H}&=(\mathbb{T}\boldsymbol{E})^{\mathsf{T}}\mathbb{T}^{\mathsf{T}}\mathbb{T}(\mathbb{T}\boldsymbol{H})\\&=((\boldsymbol{E}\times\boldsymbol{n})\times\boldsymbol{n})\cdot((\boldsymbol{H}\times\boldsymbol{n})\times\boldsymbol{n}).\end{aligned}$$

But the vector triple product identity, $(\boldsymbol{a}\times\boldsymbol{b})\times\boldsymbol{c}=(\boldsymbol{a}\cdot\boldsymbol{c})\boldsymbol{b}-(\boldsymbol{b}\cdot\boldsymbol{c})\boldsymbol{a}$, shows that

$$(\boldsymbol{E}\times\boldsymbol{n})\times\boldsymbol{n}=(\boldsymbol{E}\cdot\boldsymbol{n})\boldsymbol{n}-\boldsymbol{E}.$$

Hence, we have

$$\begin{aligned}\boldsymbol{E}^{\mathsf{T}}\mathbb{T}^{\mathsf{T}}\mathbb{T}\mathbb{T}^{\mathsf{T}}\mathbb{T}\boldsymbol{H}&=((\boldsymbol{E}\cdot\boldsymbol{n})\boldsymbol{n}-\boldsymbol{E})\cdot((\boldsymbol{H}\times\boldsymbol{n})\times\boldsymbol{n})\\&=-\boldsymbol{E}\cdot((\boldsymbol{H}\times\boldsymbol{n})\times\boldsymbol{n})=(\boldsymbol{E}\times\boldsymbol{n})\cdot(\boldsymbol{H}\times\boldsymbol{n})\\&=(\mathbb{T}\boldsymbol{E})\cdot(\mathbb{T}\boldsymbol{H})=\boldsymbol{E}^{\mathsf{T}}\mathbb{T}^{\mathsf{T}}\mathbb{T}\boldsymbol{H}.\end{aligned}$$

This proves that $\mathbb{T}^{\mathsf{T}}\mathbb{T}\mathbb{T}^{\mathsf{T}}\mathbb{T}=\mathbb{T}^{\mathsf{T}}\mathbb{T}$.

Exercise 60.5 (Orthogonal subscales). (i) See Step (1) in the proof of Lemma 60.9.
(ii) Setting $\eta:=u-\mathcal{I}_h^{\mathrm{b}}(u)$ in the proof of Lemma 60.10, we still obtain

$$\begin{aligned}\langle\delta_h(\mathcal{I}_h^{\mathrm{b}}(u)),w_h\rangle_{V_h',V_h}&=(\eta,\tilde{A}_h(w_h))_L+\frac{1}{2}((\mathcal{M}^{\mathrm{BP}}+\mathcal{N})\eta,w_h)_{L(\partial D)}\\&\quad+\overline{n_h(w_h,\eta)}+\sum_{F\in\mathcal{F}_h^\circ}(\mathcal{S}_F^\circ[\![\eta]\!],[\![w_h]\!])_{L(F)}\\&=:\mathfrak{T}_1+\mathfrak{T}_2+\mathfrak{T}_3+\mathfrak{T}_4.\end{aligned}$$

The only difficulty lies in bounding $\mathfrak{T}_1$ since we have not included the term $\|\tau^{\frac12}A_{1h}(w_h)\|_L$ in the $\|\cdot\|_{V_h}$-norm. Since $(\eta,\underline{A}_{1h}(w_h))_L=0$ by definition of $\mathcal{I}_h^{\mathrm{b}}$, we infer that

$$|(\eta,A_{1h}(w_h))_L|=|(\eta,(A_{1h}-\underline{A}_{1h})(w_h))_L|\le\mu_0^{\frac12}\|\eta\|_L\mu_0^{-\frac12}L_{\mathcal{A}}\|w_h\|_L,$$

where we used the fact that the fields $\mathcal{A}^k$ are Lipschitz and an inverse inequality to estimate $\|\nabla w_h\|_{\boldsymbol{L}(K)}$. Since $L_{\mathcal{A}}\le c\mu_0$, this gives the expected bound on $|\langle\delta_h(\mathcal{I}_h^{\mathrm{b}}(u)),w_h\rangle_{V_h',V_h}|$.
(iii) Adapting the proof of Lemma 27.8 where we bound the infimum over $v_h\in V_h$ by taking $v_h:=\mathcal{I}_h^{\mathrm{b}}(u)$, and using the stability property from Step (i) together with the consistency/boundedness property from Step (ii), we infer that

$$\|u-u_h\|_{V_\flat}\le c\,\|u-\mathcal{I}_h^{\mathrm{b}}u\|_{V_\sharp}.$$

Finally, we use the approximation properties of $\mathcal{I}_h^{\mathrm{b}}$ to derive the error estimate.

Chapter 61

Advection-diffusion

Exercises

Exercise 61.1 (A priori estimates). Consider the problem (61.1). Assume that $\mathbb{d}_\epsilon := \epsilon \mathbb{I}_d$, $\nabla\cdot\boldsymbol{\beta} = 0$, $\boldsymbol{\beta}_{|\partial D} = \mathbf{0}$, $\mu := \mu_0 \geq 0$, and $f \in H^1_0(D)$. Let $\nabla_{\mathrm{s}}\boldsymbol{\beta} := \frac{1}{2}(\nabla\boldsymbol{\beta} + (\nabla\boldsymbol{\beta})^\intercal)$ denote the symmetric part of the gradient of $\boldsymbol{\beta}$, and assume that there is $\mu_0' > 0$ s.t. $\nabla_{\mathrm{s}}\boldsymbol{\beta} + \mu\mathbb{I}_d \geq \mu_0'\mathbb{I}_d$ in the sense of quadratic forms. Prove that $|u|_{H^1(D)} \leq (\mu_0' + \mu_0)^{-1}|f|_{H^1(D)}$ and $\|\Delta u\|_{L^2(D)} \leq (4(\mu_0' + \mu_0)\epsilon)^{-\frac{1}{2}}|f|_{H^1(D)}$. (*Hint*: test the PDE (61.1) with $-\Delta u$.) *Note*: see also Beirão da Veiga [3], Burman [8].

Exercise 61.2 (Advection-diffusion, 1D). Let $D := (0,1)$ and let ϵ, b be two positive real numbers. Let $f : D \to \mathbb{R}$ be a smooth function. Consider the PDE $-\epsilon u'' + bu' = f$ in D with the boundary conditions $u(0) = 0$, $u(1) = 0$. Consider H^1-conforming $\mathbb{P}_1$ Lagrange finite elements on the uniform grid $\mathcal{T}_h$ with nodes $x_i := ih$, $\forall i \in \{0{:}I\}$, and meshsize $h := \frac{1}{I+1}$. (i) Evaluate the stiffness matrix. (*Hint*: factor out the ratio $\frac{\epsilon}{h}$ and introduce the local Péclet number $\gamma := \frac{bh}{\epsilon}$.) (ii) Solve the linear system when $f := 1$ and plot the solutions for $h := 10^{-2}$ and $\gamma \in \{0.1, 1, 10\}$. (*Hint*: the solution $\mathsf{U} \in \mathbb{R}^I$ has the form $\mathsf{U}^0 + \tilde{\mathsf{U}}$ with $\mathsf{U}^0_i := b^{-1}ih$ and $\tilde{\mathsf{U}}_i := \varrho + \theta\delta^i$ for some constants ϱ, θ, δ.) (iii) Consider now the boundary conditions $u(0) = 0$ and $u'(1) = 0$. Write the weak formulation and show well-posedness. Evaluate the stiffness matrix. (*Hint*: this matrix is now of order $(I+1)$.) Derive the equation satisfied by $h^{-1}(\mathsf{U}_{I+1} - \mathsf{U}_I)$, and comment on the limit values obtained as $h \to 0$ with fixed $\epsilon > 0$ and as $\epsilon \to 0$ with fixed $h \in \mathcal{H}$.

Exercise 61.3 (Artificial viscosity). Consider the model problem (61.1) with $\mathbb{d} := \epsilon\mathbb{I}_d$ with constant $\epsilon > 0$. Assume that $u \in H^2(D)$. Assume that $\boldsymbol{\beta}$ is divergence-free and $\mu > 0$ is constant, and set $b := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$. Consider the finite element space $V_h := P^{\mathrm{g}}_{1,0}(\mathcal{T}_h)$ on a mesh from a quasi-uniform sequence (for simplicity). Consider the following nonconsistent approximation: Find $u_h \in V_h$ such that $a_\epsilon(u_h, w_h) + s_h(u_h, w_h) = (f, w_h)_{L^2(D)}$ for all $w_h \in V_h$, where $s_h(v_h, w_h) := \frac{1}{2}bh(\nabla v_h, \nabla w_h)_{\boldsymbol{L}^2(D)}$ for all $v_h, w_h \in P^{\mathrm{g}}_{1,0}(\mathcal{T}_h)$. (i) Prove the following error estimate:

$$\mu^{\frac{1}{2}}\|u - u_h\|_{L^2(D)} + (\epsilon^{\frac{1}{2}} + (bh)^{\frac{1}{2}})\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \leq c\,(\epsilon^{\frac{1}{2}} + (bh)^{\frac{1}{2}} + \mu^{\frac{1}{2}}h + \mu^{-\frac{1}{2}}b)h|u|_{H^2(D)}.$$

(*Hint*: use the norms $\|v\|^2_{V_\flat} := (\epsilon + \frac{1}{2}bh)\|\nabla v\|^2_{\boldsymbol{L}^2(D)} + \mu\|v\|^2_{L^2(D)}$, $\|v\|^2_{V_\sharp} := (\epsilon + \frac{1}{2}bh)\|\nabla v\|^2_{\boldsymbol{L}^2(D)} + (\mu + 2bh^{-1})\|v\|^2_{L^2(D)}$ and adapt the proof of Lemma 27.8.) (ii) Consider the 1D setting of Exercise 61.2 with $f := 1$. Set $V_h := P^{\mathrm{g}}_{1,0}(\mathcal{T}_h) = \mathrm{span}\{\varphi_i\}_{i\in\{1:I\}}$, where the φ_i's are the usual hat basis functions

in $P^{\mathrm{g}}_{1,0}(\mathcal{T}_h)$. Let $\xi : [0,1] \to \mathbb{R}$ be a smooth function, called bubble function, s.t. $\xi(0) = \xi(1) = 0$ and $\xi \geq 0$. For all $i \in \{1{:}I\}$, set $\xi_i(x) := \xi(\frac{x-x_{i-1}}{h})$ if $x \in [x_{i-1}, x_i]$, $\xi_i(x) := -\xi(\frac{x-x_i}{h})$ if $x \in [x_i, x_{i+1}]$, and $\xi_i(x) := 0$ otherwise, and set $\psi_i := \varphi_i + \xi_i$. Let $W_h = \mathrm{span}\{\psi_i\}_{i\in\{1:I\}}$. Prove that the Petrov–Galerkin formulation using the pair (V_h, W_h) as trial and test spaces is equivalent to a Galerkin formulation in V_h with the bilinear form augmented by an artificial viscosity term. (*Hint*: verify that $\int_{x_{i-1}}^{x_{i+1}} u_h' \xi_i \,\mathrm{d}x = h(\int_0^1 \xi(x)\,\mathrm{d}x) \int_{x_{i-1}}^{x_{i+1}} u_h' \varphi_i' \,\mathrm{d}x$ for all $i \in \{1{:}I\}$.) Explain how to choose $\int_0^1 \xi(x)\,\mathrm{d}x$ so that the stiffness matrix is always an M-matrix. (*Hint*: use Exercise 61.2.)

Exercise 61.4 (Bound on consistency term). Prove Lemma 61.7. (*Hint*: observe that $|\boldsymbol{n}\cdot\mathrm{d}_\epsilon \nabla v_h| \leq \lambda_F^{\frac12} \|\mathrm{d}_\epsilon^{\frac12} \nabla v_h\|_{\ell^2(\mathbb{R}^d)}$, use that $\mathrm{d}_\epsilon^{\frac12}\nabla v_h$ is a piecewise polynomial, and adapt the proof of Lemma 37.2.)

Exercise 61.5 (Divergence-free advection). (i) Prove (61.27). (*Hint*: use Lemma 22.3 and $[\![\zeta_0 v_h]\!] = [\![\zeta_0]\!] v_h$, and bound $[\![\zeta_0]\!]$ using L_ζ.) (ii) Prove (61.28). (*Hint*: use that $\|\varphi_h - \zeta v_h\|_{L^2(K)} \leq \|\varphi_h - \zeta v_h\|_{L^2(K)} + \|(\zeta - \zeta_0) v_h\|_{L^2(K)}$.) (iii) Prove that $\|\varphi_h\|_{V_h} \leq c \|v_h\|_{V_h}$. (*Hint*: bound $\|\zeta_0 v_h\|_{V_h}$ and $\|\varphi_h - \zeta_0 v_h\|_{V_h}$.)

Solution to exercises

Exercise 61.1 (A priori estimates). Following the hint and integrating by parts, we infer that

$$\epsilon \|\Delta u\|^2_{L^2(D)} - (\boldsymbol{\beta}\cdot\nabla u, \Delta u)_{L^2(D)} + \mu_0 |u|^2_{H^1(D)} = -(f, \Delta u)_{L^2(D)} = (\nabla f, \nabla u)_{\boldsymbol{L}^2(D)},$$

where we used that $\mathrm{d}_\epsilon = \epsilon \mathbb{I}_d$, $u \in H^1_0(D)$, $\mu := \mu_0$, and $f \in H^1_0(D)$. Using that $\boldsymbol{\beta}_{|\partial D} = \boldsymbol{0}$, we also infer that

$$\begin{aligned}
-(\boldsymbol{\beta}\cdot\nabla u, \Delta u)_{L^2(D)} &= -\sum_{i,j\in\{1:d\}} (\beta_i \partial_i u, \partial_j \partial_j u)_{L^2(D)} \\
&= \sum_{i,j\in\{1:d\}} ((\partial_j \beta_i)\partial_i u, \partial_j u)_{L^2(D)} + (\beta_i \partial_i(\partial_j u), \partial_j u)_{L^2(D)} \\
&=: \mathfrak{T}_1 + \mathfrak{T}_2.
\end{aligned}$$

We have $\mathfrak{T}_1 = ((\nabla_{\mathrm{s}}\boldsymbol{\beta})\nabla u, \nabla u)_{\boldsymbol{L}^2(D)}$. Using that $\nabla\cdot\boldsymbol{\beta} = 0$ and using again that $\boldsymbol{\beta}$ vanishes at the boundary, we obtain that

$$\mathfrak{T}_2 = \sum_{i,j\in\{1:d\}} (\boldsymbol{\beta}\cdot\nabla\partial_j u, \partial_j u)_{L^2(D)} = \int_D \frac12 \nabla\cdot(\boldsymbol{\beta}\|\nabla u\|^2)\,\mathrm{d}x = 0.$$

In summary, we have shown that

$$\epsilon \|\Delta u\|^2_{L^2(D)} + ((\nabla_{\mathrm{s}}\boldsymbol{\beta})\nabla u, \nabla u)_{\boldsymbol{L}^2(D)} + \mu_0 |u|^2_{H^1(D)} = (\nabla f, \nabla u)_{\boldsymbol{L}^2(D)}.$$

Our assumption on $\nabla_{\mathrm{s}}\boldsymbol{\beta}$ implies that

$$\epsilon \|\Delta u\|^2_{L^2(D)} + (\mu_0' + \mu_0) |u|^2_{H^1(D)} \leq (\nabla f, \nabla u)_{\boldsymbol{L}^2(D)}.$$

The assertion follows by bounding the right-hand side as in the proof of the a priori estimate (61.5).

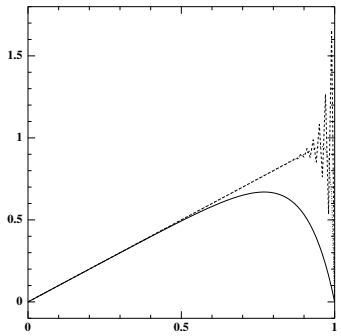
Exercise 61.2 (Advection-diffusion, 1D). (i) The stiffness matrix is given by $\mathcal{A} = \frac{\epsilon}{h}\text{tridiag}(-1-\frac{\gamma}{2}, 2, -1+\frac{\gamma}{2})$.
(ii) Assuming that $f := 1$, the linear system to be solved is $\mathcal{A}\mathsf{U} = h(1,\dots,1)^\mathsf{T}$. Since $\mathcal{A}\mathsf{U}^0 = (h,\dots,h,h+\gamma^{-1}(1-\frac{\gamma}{2}))^\mathsf{T}$ (observe that $h(I+1)=1$), we infer that $\mathcal{A}\tilde{\mathsf{U}} = (0,\dots,0,\gamma^{-1}(\frac{\gamma}{2}-1))^\mathsf{T}$. If $\gamma = 2$, then $\tilde{\mathsf{U}} = 0$. Let us assume now that $\gamma \neq 2$. Using $\tilde{\mathsf{U}}_i = \varrho + \theta\delta^i$, we infer from the rows $\{2{:}I-1\}$ of the linear system that

$$\left(-1-\tfrac{\gamma}{2}\right) + 2\delta + \left(-1+\tfrac{\gamma}{2}\right)\delta^2 = 0,$$

so that $\delta = 1$ or $\delta = \frac{2+\gamma}{2-\gamma}$. The first row of the system yields $\theta = -\varrho$. From the last row of the system, we finally infer that $\frac{\epsilon}{h}(1-\frac{\gamma}{2})\varrho(1-\delta^{I+1}) = \gamma^{-1}(\frac{\gamma}{2}-1)$, i.e., $b\varrho(1-\delta^{I+1}) = -1$. Notice that $\delta \neq 1$ because we assumed $\gamma = \frac{bh}{\epsilon} \neq 0$. Hence, $-\theta = \varrho = -b^{-1}(1-\delta^{I+1})^{-1}$, that is,

$$\tilde{\mathsf{U}}_i = -b^{-1}\frac{\delta^i - 1}{\delta^{I+1}-1}, \qquad \delta = \frac{2+\gamma}{2-\gamma}.$$

When $\gamma > 2$, the components of the vector $\tilde{\mathsf{U}}$ oscillate between positive and negative values. The approximate solutions for $\gamma \in \{0.1, 1, 10\}$ obtained with $h := 10^{-2}$ are plotted on the figure shown here. We observe that for $\gamma = 10$ the approximate solution exhibits spurious oscillations close to the boundary layer. Instead, the approximate solutions for $\gamma = 1$ and $\gamma = 0.1$ match well the exact solution.



(iii) Setting $V := \{v \in H^1(D) \mid v(0) = 0\}$, the weak formulation now consists of seeking $u \in V$ such that $a(u,w) = \ell(w)$ for all $w \in V$. Since $\int_0^1 bv'v\,\mathrm{d}x = \frac{1}{2}bv(1)^2 \geq 0$, the bilinear form a is still coercive on V. The stiffness matrix is of order $(I+1)$ and has the following tridiagonal structure:

$$\mathcal{A} = \frac{\epsilon}{h}\begin{pmatrix} c_0 & c_+ & 0 & \dots & 0 \\ c_- & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & c_0 & c_+ \\ 0 & \dots & 0 & c_- & c_0' \end{pmatrix}$$

with $c_0 := 2$, $c_0' := 1+\frac{\gamma}{2}$, $c_+ := -1+\frac{\gamma}{2}$, and $c_- := -1-\frac{\gamma}{2}$. We infer that $(\epsilon + \frac{bh}{2})(\mathsf{U}_{I+1} - \mathsf{U}_I) = \int_{x_I}^{x_{I+1}} f\varphi_{I+1}\,\mathrm{d}x$, so that

$$\frac{\mathsf{U}_{I+1}-\mathsf{U}_I}{h} = \frac{2\int_{x_I}^{x_{I+1}} f\varphi_{I+1}\,\mathrm{d}x}{2\epsilon + bh}.$$

Hence, $\frac{\mathsf{U}_{I+1}-\mathsf{U}_I}{h} \to 0$ as $h \to 0$ with fixed $\epsilon > 0$, whereas $\frac{\mathsf{U}_{I+1}-\mathsf{U}_I}{h} \to \frac{f(1)}{b}$ as $\epsilon \to 0$ with fixed $h \in \mathcal{H}$.

Exercise 61.3 (Artificial viscosity). (i) Let us introduce the following stability norm:

$$\|v_h\|_{V_h}^2 := (\epsilon + \tfrac{1}{2}bh)\|\nabla v_h\|_{\boldsymbol{L}^2(D)}^2 + \mu\|v_h\|_{L^2(D)}^2.$$

The coercivity of the discrete bilinear form $a_\epsilon + s_h$ on V_h in this norm is straightforward (with coercivity constant $\alpha := 1$). Let us set $V_\mathrm{s} := H^2(D) \cap H^1_0(D)$, $V_\sharp := V_\mathrm{s} + V_h$, and let us equip the space $V_\sharp$ with the following norms:

$$\|v\|^2_{V_\flat} := (\epsilon + \tfrac{1}{2}bh)\|\nabla v\|^2_{\boldsymbol{L}^2(D)} + \mu\|v\|^2_{L^2(D)},$$
$$\|v\|^2_{V_\sharp} := (\epsilon + \tfrac{1}{2}bh)\|\nabla v\|^2_{\boldsymbol{L}^2(D)} + (\mu + 2bh^{-1})\|v\|^2_{L^2(D)}.$$

Notice that (27.7) is satisfied with $c_\flat := 1$ (i.e., $\|v_h\|_{V_\flat} \le \|v_h\|_{V_h}$ on V_h and $\|v\|_{V_\flat} \le \|v\|_{V_\sharp}$ on $V_\sharp$). Recalling Definition 27.3, the consistency error is such that for all $v_h, w_h \in V_h$,

$$\begin{aligned}\langle \delta_h(v_h), w_h\rangle_{V_h',V_h} &= (f, w_h)_{L^2(D)} - a_\epsilon(v_h, w_h) - s_h(v_h, w_h)\\ &= a_\epsilon(\eta, w_h) + s_h(\eta, w_h) - s_h(u, w_h),\end{aligned}$$

with $\eta := u - v_h$. Integrating by parts the advective derivative, we infer that

$$(\boldsymbol{\beta}\cdot\nabla v, w_h)_{L^2(D)} \le \|v\|_{L^2(D)} b\|\nabla w_h\|_{\boldsymbol{L}^2(D)} \le (2b)^{\frac{1}{2}} h^{-\frac{1}{2}} \|v\|_{L^2(D)} s_h(w_h, w_h)^{\frac{1}{2}}.$$

This implies that

$$|a_\epsilon(\eta, w_h)| \le \|\eta\|_{V_\sharp}\|w_h\|_{V_h}.$$

Moreover, we have

$$|s_h(\eta, w_h)| \le s_h(\eta,\eta)^{\frac{1}{2}} s_h(v_h, v_h)^{\frac{1}{2}} \le \|\eta\|_{V_\sharp}\|w_h\|_{V_h}.$$

Integrating by parts, using the Cauchy–Schwarz inequality and the definition of the stability norm $\|\cdot\|_{V_h}$, we finally have

$$\begin{aligned}|s_h(u, w_h)| = \frac{1}{2}bh|(\Delta u, w_h)_{L^2(D)}| &\le \frac{1}{2}bh\|\Delta u\|_{L^2(D)}\|w_h\|_{L^2(D)}\\ &\le \frac{1}{2}\mu^{-\frac{1}{2}}bh\|\Delta u\|_{L^2(D)}\|w_h\|_{V_h}.\end{aligned}$$

Putting the above bounds together, we infer that

$$\|\delta_h(v_h)\|_{V_h'} \le c\,\|u - v_h\|_{V_\sharp} + \frac{1}{2}\mu^{-\frac{1}{2}}bh\|\Delta u\|_{L^2(D)}.$$

Adapting the proof of Lemma 27.8, we obtain

$$\|u - u_h\|_{V_\flat} \le c\left(\inf_{v_h\in V_h}\|u - v_h\|_{V_\sharp} + \frac{1}{2}\mu^{-\frac{1}{2}}bh\|\Delta u\|_{L^2(D)}\right).$$

Using the approximation capacity of the discrete space $V_h = P^\mathrm{g}_{1,0}(\mathcal{T}_h)$, we infer that

$$\inf_{v_h\in V_h}\|u - v_h\|_{V_\sharp} \le c\,(\epsilon^{\frac{1}{2}} + (bh)^{\frac{1}{2}} + \mu^{\frac{1}{2}}h)h|u|_{H^2(D)}.$$

Since $\|\Delta u\|_{L^2(D)} \le c\,|u|_{H^2(D)}$, this leads to the expected error bound.
(ii) The Petrov–Galerkin approximation consists of seeking $u_h \in V_h$ such that

$$\int_0^1 \epsilon u_h'(\varphi_i' + \xi_i')\,\mathrm{d}x + \int_0^1 bu_h'(\varphi_i + \xi_i)\,\mathrm{d}x = \int_0^1 f(\varphi_i + \xi_i)\,\mathrm{d}x,$$

for all $i \in \{1{:}I\}$. Since u_h' is piecewise constant, we infer that $\int_0^1 \epsilon u_h'\xi_i'\,\mathrm{d}x = 0$. Moreover, a direct calculation using again that u_h' is piecewise constant shows that

$$\int_{x_{i-1}}^{x_{i+1}} u_h'\xi_i\,\mathrm{d}x = h\Big(\int_0^1 \xi(x)\,\mathrm{d}x)\Big)\int_{x_{i-1}}^{x_{i+1}} u_h'\varphi_i'\,\mathrm{d}x.$$

Since $f := 1$ and $\int_0^1 \xi_i \,\mathrm{d}x = 0$, we finally infer that $\int_0^1 f\xi_i \,\mathrm{d}x = 0$. Hence, $u_h \in V_h$ is such that

$$\int_0^1 (\epsilon + \epsilon_h) u_h' \varphi_i' \,\mathrm{d}x + \int_0^1 b u_h' \varphi_i \,\mathrm{d}x = \int_0^1 f\varphi_i \,\mathrm{d}x,$$

where $\epsilon_h = bh \int_0^1 b(x)\,\mathrm{d}x$ is an artificial viscosity. To obtain an M-matrix, the condition derived in Exercise 61.2 is $\gamma_h \le 2$ with local Péclet number $\gamma_h = \frac{bh}{\epsilon+\epsilon_h}$, for which a sufficient condition is $\int_0^1 \xi(x)\,\mathrm{d}x \ge \frac{1}{2}$.

Exercise 61.4 (Bound on consistency term). Let $v_h, w_h \in V_h$. Let $F \in \mathcal{F}_h^\partial$. Since $\mathbb{d}_\epsilon$ is symmetric positive definite, we have

$$|\boldsymbol{n}\cdot\mathbb{d}_\epsilon \nabla v_h| \le (\boldsymbol{n}\cdot\mathbb{d}_\epsilon\boldsymbol{n})^{\frac12}(\nabla v_h\cdot\mathbb{d}_\epsilon\nabla v_h)^{\frac12} = \lambda_F^{\frac12}\|\mathbb{d}_\epsilon^{\frac12}\nabla v_h\|_{\ell^2(\mathbb{R}^d)}.$$

Using the discrete trace inequality $h_F^{\frac12}\|\mathbb{d}_\epsilon^{\frac12}\nabla v_h\|_{\boldsymbol{L}^2(F)} \le c_{\mathrm{dt}}\|\mathbb{d}_\epsilon^{\frac12}\nabla v_h\|_{\boldsymbol{L}^2(K_l)}$ (this is legitimate since $\mathbb{d}_\epsilon^{\frac12}\nabla v_h$ is a piecewise polynomial because $\mathbb{d}_\epsilon$ is piecewise constant) and the Cauchy–Schwarz inequality, we infer that

$$\begin{aligned}\left|\int_{\partial D} (\boldsymbol{n}\cdot\mathbb{d}_\epsilon\nabla v_h) w_h \,\mathrm{d}s\right| &\le \Bigg(\sum_{F\in\mathcal{F}_h^\partial} h_F\|\mathbb{d}_\epsilon^{\frac12}\nabla v_h\|_{\boldsymbol{L}^2(F)}^2\Bigg)^{\frac12}\Bigg(\sum_{F\in\mathcal{F}_h^\partial}\frac{\lambda_F}{h_F}\|w_h\|_{L^2(F)}^2\Bigg)^{\frac12}\\ &\le c_{\mathrm{dt}}\Bigg(\sum_{F\in\mathcal{F}_h^\partial}\|\mathbb{d}_\epsilon^{\frac12}\nabla v_h\|_{\boldsymbol{L}^2(K_l)}^2\Bigg)^{\frac12}\Bigg(\sum_{F\in\mathcal{F}_h^\partial}\frac{\lambda_F}{h_F}\|w_h\|_{L^2(F)}^2\Bigg)^{\frac12}.\end{aligned}$$

The assertion follows by rewriting the summation over $F \in \mathcal{F}_h^\partial$ as a summation over $K \in \mathcal{T}_h^{\partial D}$ and by using the definition of n_∂.

Exercise 61.5 (Divergence-free advection). (i) Proof of (61.27). The Cauchy–Schwarz inequality, together with discrete trace and inverse inequalities and $\varpi_0 \ge 1$ show that

$$\begin{aligned}|a_1(v_h, \varphi_h - \zeta_0 v_h)| \le c\,\varpi_0^{\frac12}(A_1+A_2+A_3)\times\Bigg(&\sum_{K\in\mathcal{T}_h}\lambda_{\sharp,K}h_K^{-2}\|\varphi_h-\zeta_0 v_h\|_{L^2(K)}^2\\ &+\tau_K\delta_K(\beta_K^2h_K^{-2}+\lambda_{\sharp,K}^2h_K^{-4})\|\varphi_h-\zeta_0v_h\|_{L^2(K)}^2\Bigg)^{\frac12}.\end{aligned}$$

We observe that

$$\tau_K\delta_K(\beta_K^2h_K^{-2}+\lambda_{\sharp,K}^2h_K^{-4}) \le \beta_K h_K^{-1} + \lambda_{\sharp,K}h_K^{-2},$$

where the bound on the first term follows from $\delta_K \le 1$ and the bound on the second term follows from $\delta_K \le \rho_K^{-1}\mathrm{Pe}_K$ with $\mathrm{Pe}_K := \frac{h_K^2}{\tau_K\lambda_{\flat,K}}$. As a result, we obtain

$$|a_1(v_h, \varphi_h - \zeta_0 v_h)| \le c\,\varpi_0^{\frac12}(A_1+A_2+A_3)\times\Bigg(\sum_{K\in\mathcal{T}_h}(\beta_K h_K^{-1}+\lambda_{\sharp,K}h_K^{-2})\|\varphi_h-\zeta_0v_h\|_{L^2(K)}^2\Bigg)^{\frac12}.$$

Invoking Lemma 22.3 and observing that $[\![\zeta_0 v_h]\!] = [\![\zeta_0]\!]v_h$, $[\![\zeta_0]\!] \le ch_K L_\zeta$, we infer that

$$\|\varphi_h-\zeta_0v_h\|_{L^2(K)}^2 \le c\,h_K\sum_{F\in\check{\mathcal{F}}_K^\circ}h_K^2L_\zeta^2\|v_h\|_{L^2(F)}^2 \le c'\,h_K^2L_\zeta^2\|v_h\|_{L^2(K)}^2,$$

where we used a discrete trace inequality and the regularity of the mesh sequence (recall that $\check{\mathcal{F}}_K^\circ$ is the collection of the mesh interfaces sharing at least a vertex with K). Using the assumption $L_\zeta^2 \max(\lambda_{\sharp,K}, \beta_K h_K) \le \mu_0$, we conclude that

$$|a_1(v_h, \varphi_h - \zeta_0 v_h)| \le c\, \varpi_0^{\frac12} (A_1 + A_2 + A_3) \mu_0^{\frac12} \|v_h\|_{L^2(D)},$$

which leads to the expected bound.
(ii) Proof of (61.28). The Cauchy–Schwarz inequality and the triangle inequality lead to

$$|a_2(v_h, \varphi_h - \zeta v_h)| \le c \left(A_3^2 + A_4^2 + \sum_{K\in\mathcal{T}_h} \tau_K \delta_K \|\nabla\cdot(\mathrm{d}_\epsilon \nabla v_h)\|_{L^2(K)}^2 \right)^{\frac12} \times \left(\sum_{K\in\mathcal{T}_h} \tau_K^{-1}\delta_K^{-1} \|\varphi_h - \zeta v_h\|_{L^2(K)}^2 + \sum_{F\in\mathcal{F}_h^\partial} \beta_{K_l} h_{K_l}^{-1} \|\varphi_h - \zeta v_h\|_{L^2(K_l)}^2 \right)^{\frac12}.$$

Using that $\tau_K \delta_K \|\nabla\cdot(\mathrm{d}_\epsilon \nabla v_h)\|_{L^2(K)}^2 \le c \|\mathrm{d}_\epsilon^{\frac12} \nabla v_h\|_{\boldsymbol{L}^2(K)}^2$ since $\tau_K \delta_K \lambda_{\sharp,K} h_K^{-2} \le 1$, we obtain after rearranging some terms that

$$|a_2(v_h, \varphi_h - \zeta v_h)| \le c\,(A_1 + A_3 + A_4) \times \left(\sum_{K\in\mathcal{T}_h} \max(\tau_K^{-1}\delta_K^{-1}, \beta_K h_K^{-1}) \|\varphi_h - \zeta v_h\|_{L^2(K)}^2 \right)^{\frac12}.$$

Using the triangle inequality $\|\varphi_h - \zeta v_h\|_{L^2(K)} \le \|\varphi_h - \zeta_0 v_h\|_{L^2(K)} + \|(\zeta - \zeta_0) v_h\|_{L^2(K)}$, we bound the first term using as above and the second one using $\|\zeta - \zeta_0\|_{L^\infty(K)} \le c h_K L_\zeta$. This yields

$$|a_2(v_h, \varphi_h - \zeta v_h)| \le c\,(A_1 + A_3 + A_4) \times \left(\sum_{K\in\mathcal{T}_h} \max(\tau_K^{-1}\delta_K^{-1} h_K^2, \beta_K h_K) L_\zeta^2 \|v_h\|_{L^2(K)}^2 \right)^{\frac12}.$$

To prove the expected bound, it remains to verify that

$$\max(\tau_K^{-1}\delta_K^{-1} h_K^2, \beta_K h_K) L_\zeta^2 \le \mu_0.$$

The bound $\beta_K h_K L_\zeta^2 \le \mu_0$ follows from the assumption on L_ζ. Let us consider $\tau_K^{-1}\delta_K^{-1} h_K^2 L_\zeta^2$. Using the definition of δ_K, we obtain

$$\tau_K^{-1}\delta_K^{-1} h_K^2 L_\zeta^2 = \max(\tau_K^{-1} h_K^2 L_\zeta^2, \lambda_{\sharp,K} L_\zeta^2).$$

The second argument verifies $\lambda_{\sharp,K} L_\zeta^2 \le \mu_0$ by assumption on L_ζ. Using the definition of τ_K, we finally have

$$\tau_K^{-1} h_K^2 L_\zeta^2 = \max(\beta_K h_K^{-1}, \mu_0) h_K^2 L_\zeta^2 = \max(\beta_K h_K L_\zeta^2, \mu_0 h_K^2 L_\zeta^2) \le \mu_0,$$

since we assumed that $\beta_K h_K L_\zeta^2 \le \mu_0$ and $h_K L_\zeta \le 1$.
(iii) Let us prove that $\|\varphi_h\|_{V_h} \le c \|v_h\|_{V_h}$. The triangle inequality yields $\|\varphi_h\|_{V_h} \le \|\zeta_0 v_h\|_{V_h} + \|\varphi_h - \zeta_0 v_h\|_{V_h}$, and since ζ_0 is piecewise constant, we have $\|\zeta_0 v_h\|_{V_h} \le \zeta_\sharp \|v_h\|_{V_h}$. Using inverse inequalities, we finally infer that

$$\|\varphi_h - \zeta_0 v_h\|_{V_h}^2 \le \sum_{K\in\mathcal{T}_h} \Big(\lambda_{\sharp,K} h_K^{-2} + \beta_K h_K^{-1} + \mu_0 \Big) \|\varphi_h - \zeta_0 v_h\|_{L^2(K)}^2,$$

and we conclude as above.

Chapter 62

Stokes equations: Residual-based stabilization

Exercises

Exercise 62.1 (Pressure gradient). Assume (62.14). Prove an inf-sup condition similar to (62.14) using the norm $\|(\boldsymbol{v}_h, q_h)\|_{Y_h^+}^2 := \|(\boldsymbol{v}_h, q_h)\|_{Y_h}^2 + \sum_{K\in\mathcal{T}_h} \mu^{-1} h_K^2 \|\nabla q_h\|_{\boldsymbol{L}^2(K)}^2$. (*Hint*: use an inverse inequality.)

Exercise 62.2 (Inf-sup partner). The objective of this exercise is to reprove the inf-sup condition (62.14) by identifying an inf-sup partner for all $(v_h, q_h) \in Y_h$ as suggested in Remark 25.10. (i) Prove that there is $\rho \in (0,1)$ s.t. $t_h((\boldsymbol{v}_h, q_h), ((1-\rho)\boldsymbol{v}_h + \rho\boldsymbol{w}_h, (1-\rho)q_h)) \geq \eta \|(\boldsymbol{v}_h, q_h)\|_{Y_h}^2$ with $\boldsymbol{w}_h := \boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{w}_{q_h})$ and $\boldsymbol{w}_{q_h}$ defined in (62.16). (*Hint*: use (62.15) and the bounds on $\mathfrak{T}_2, \mathfrak{T}_3$ from the proof of Lemma 62.3.) (ii) Show that the inf-sup condition (62.14) is satisfied with a constant γ_0 depending on ρ, β_D, η, and the constant c_w introduced in (62.18), i.e., $\|(\boldsymbol{w}_h, 0)\|_{Y_h} \leq c_w \mu^{\frac{1}{2}} |\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)}$. (*Hint*: identify an appropriate inf-sup partner for (v_h, q_h) and use Remark 25.10.)

Exercise 62.3 (Approximation). Let $|\cdot|_S$ be the GaLS stabilization seminorm, i.e., $|\cdot|_S^2 = |\cdot|_{S^{\mathrm{r}}}^2 + |\cdot|_{S^{\mathrm{p}}}^2 + |\cdot|_{S^{\mathrm{n}}}^2$. Let $(\boldsymbol{\eta}, \zeta) \in (\boldsymbol{H}^2(D) \times H^1(D)) \cap Y$ be s.t. $\mathbb{r}(\boldsymbol{\eta}, \zeta)_{|\partial D_{\mathrm{n}}} \boldsymbol{n} = \boldsymbol{0}$. (i) Prove that $|(\boldsymbol{\eta}, \zeta)|_S \leq ch(\mu^{\frac{1}{2}} |\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)} + \mu^{-\frac{1}{2}} |\zeta|_{H^1(D)})$. (ii) Prove that $|(\boldsymbol{\eta} - \boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{\eta}), \zeta - \mathcal{I}_h^{\mathrm{p}}(\zeta))|_S \leq ch(\mu^{\frac{1}{2}} |\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)} + \mu^{-\frac{1}{2}} |\zeta|_{H^1(D)})$. (*Hint*: use (62.24).) (iii) Estimate $|(\boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{\eta}), \mathcal{I}_h^{\mathrm{p}}(\zeta))|_S$.

Exercise 62.4 (Inf-sup condition on t_h). Assume that $\partial D = \partial D_{\mathrm{d}}$ so that $\boldsymbol{V}_{\mathrm{d}} := \boldsymbol{H}_0^1(D)$. Reprove (62.14) by accepting as a fact (see Exercise 63.2) that there is $\beta_0 > 0$ s.t. for all $h \in \mathcal{H}$ and all $q_h \in Q_h$,

$$\beta_0 \, \mu^{-\frac{1}{2}} \|q_h\|_{L^2(D)} \leq \sup_{\boldsymbol{w}_h \in \boldsymbol{V}_{h\mathrm{d}}} \frac{|b(\boldsymbol{w}_h, q_h)|}{\mu^{\frac{1}{2}} |\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}} + |q_h|_{S^{\mathrm{gp}}} + |q_h|_{S^{\mathrm{p}}},$$

with $|q_h|_{S^{\mathrm{gp}}}^2 := \sum_{F\in\mathcal{F}_h^\circ} \frac{h_F^3}{\mu} \|[\![\nabla_h q_h]\!] \cdot \boldsymbol{n}_F\|_{L^2(F)}^2$ for all $q_h \in Q_h$. (*Hint*: use that $b(\boldsymbol{w}_h, q_h) = t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, 0)) - a(\boldsymbol{v}_h, \boldsymbol{w}_h) - s_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, 0))$ for all $\boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}$, and prove that $|q_h|_{S^{\mathrm{gp}}}^2 \leq c\big(|(\boldsymbol{v}_h, q_h)|_{S^{\mathrm{r}}}^2 + \mu |\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2\big)$.)

Solution to exercises

Exercise 62.1 (Pressure gradient). Invoking an inverse inequality, we have $h_K\|\nabla q_g\|_{\boldsymbol{L}^2(K)} \le c\|q_h\|_{L^2(K)}$ for all $K \in \mathcal{T}_h$. This implies that

$$\|(\boldsymbol{v}_h, q_h)\|_{Y_h} \le \|(\boldsymbol{v}_h, q_h)\|_{Y_h^+} \le c_+ \|(\boldsymbol{v}_h, q_h)\|_{Y_h},$$

for all $(\boldsymbol{v}_h, q_h) \in Y_h$. Therefore, we have

$$\begin{aligned} c_+^{-1}\gamma_0\|(\boldsymbol{v}_h, q_h)\|_{Y_h^+} \le \gamma_0\|(\boldsymbol{v}_h, q_h)\|_{Y_h} &\le \sup_{(\boldsymbol{w}_h, r_h)\in Y_h} \frac{t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, r_h))}{\|(\boldsymbol{w}_h, r_h)\|_{Y_h}} \\ &\le \sup_{(\boldsymbol{w}_h, r_h)\in Y_h} \frac{t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, r_h))}{\|(\boldsymbol{w}_h, r_h)\|_{Y_h^+}}, \end{aligned}$$

that is, t_h satisfies the inf-sup condition (62.14) with the constant $c_+^{-1}\gamma_0$.

Exercise 62.2 (Inf-sup partner). (i) By linearity, we have

$$\begin{aligned} t_h((\boldsymbol{v}_h, q_h), ((1-\rho)\boldsymbol{v}_h + \rho\boldsymbol{w}_h, (1-\rho)q_h)) = \\ (1-\rho)t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{v}_h, q_h)) + \rho t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, 0)). \end{aligned}$$

Using (62.15) and $1-\rho > 0$, we have

$$\begin{aligned} t_h((\boldsymbol{v}_h, q_h), ((1-\rho)\boldsymbol{v}_h + \rho\boldsymbol{w}_h, (1-\rho)q_h)) &= (1-\rho)t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{v}_h, q_h)) + \rho t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, 0)) \\ &\ge \alpha(1-\rho)\big(\mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2 + |(\boldsymbol{v}_h, q_h)|_S^2\big) \\ &\quad + \rho\big(\mu^{-1}\|q_h\|_{\boldsymbol{L}^2(D)}^2 - \mathfrak{T}_2 - \mathfrak{T}_3\big), \end{aligned}$$

where $\mathfrak{T}_2, \mathfrak{T}_3$ are defined in the proof of Lemma 62.3. Using the bounds on $\mathfrak{T}_2, \mathfrak{T}_3$ from this proof and $\rho > 0$, we infer that

$$\begin{aligned} t_h((\boldsymbol{v}_h, q_h), ((1-\rho)\boldsymbol{v}_h + \rho\boldsymbol{w}_h, (1-\rho)q_h)) &\ge \alpha(1-\rho)\big(\mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2 + |(\boldsymbol{v}_h, q_h)|_S^2\big) + \rho\mu^{-1}\|q_h\|_{\boldsymbol{L}^2(D)}^2 \\ &\quad - \rho c_{23}\big(\mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2 + |(\boldsymbol{v}_h, q_h)|_S^2\big)^{\frac12}\mu^{-\frac12}\|q_h\|_{\boldsymbol{L}^2(D)}, \end{aligned}$$

for some constant c_{23}. Applying Young's inequality leads to

$$\begin{aligned} &t_h((\boldsymbol{v}_h, q_h), ((1-\rho)\boldsymbol{v}_h + \rho\boldsymbol{w}_h, (1-\rho)q_h)) \\ &\ge \big(\alpha(1-\rho) - \tfrac12\rho c_{23}^2\big)\big(\mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2 + |(\boldsymbol{v}_h, q_h)|_S^2\big) + \tfrac12\rho\mu^{-1}\|q_h\|_{\boldsymbol{L}^2(D)}^2. \end{aligned}$$

Taking $\rho := \frac{\alpha}{2\alpha + c_{23}^2}$, we obtain $\alpha(1-\rho) - \frac12\rho c_{23}^2 = \frac12\alpha$, so that

$$t_h((\boldsymbol{v}_h, q_h), ((1-\rho)\boldsymbol{v}_h + \rho\boldsymbol{w}_h, (1-\rho)q_h)) \ge \eta\|(\boldsymbol{v}_h, q_h)\|_{Y_h}^2,$$

with $\eta := \frac12\min(\alpha, \rho)$.
(ii) The triangle inequality and $\rho \in (0,1)$ lead to

$$\|((1-\rho)\boldsymbol{v}_h + \rho\boldsymbol{w}_h, (1-\rho)q_h)\|_{Y_h} \le (1-\rho)\|(\boldsymbol{v}_h, q_h)\|_{Y_h} + \rho\|(\boldsymbol{w}_h, 0)\|_{Y_h},$$

and since

$$\|(\boldsymbol{w}_h, 0)\|_{Y_h} \le c_w\mu^{\frac12}|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)} \le c_w\beta_D^{-1}\mu^{-\frac12}\|q_h\|_{L^2(D)},$$

we infer that

$$\|((1-\rho)\boldsymbol{v}_h+\rho\boldsymbol{w}_h,(1-\rho)q_h)\|_{Y_h}\le\big((1-\rho)+\rho c_w\beta_D^{-1}\big)\|(\boldsymbol{v}_h,q_h)\|_{Y_h}.$$

Using Remark 25.10, we conclude that $((1-\rho)\boldsymbol{v}_h+\rho\boldsymbol{w}_h,(1-\rho)q_h)$ is a suitable inf-sup partner of (v_h,q_h), and the inf-sup condition (62.14) is satisfied with

$$\gamma\ge\frac{\beta_D\eta}{\beta_D(1-\rho)+\rho c_w},$$

where the value of ρ is fixed in Step (i).

Exercise 62.3 (Approximation). Let $(\boldsymbol{\eta},\zeta)\in\big(\boldsymbol{H}^2(D)\times H^1(D)\big)\cap Y$ be s.t. $\mathbb{r}(\boldsymbol{\eta},\zeta)_{|\partial D_{\rm n}}\boldsymbol{n}=\boldsymbol{0}$. Recall that $|\cdot|_S^2=|\cdot|_{S^{\rm r}}^2+|\cdot|_{S^{\rm p}}^2+|\cdot|_{S^{\rm n}}^2$.
(i) We estimate $|(\boldsymbol{\eta},\zeta)|_{S^{\rm r}}$ as follows:

$$\begin{aligned}|(\boldsymbol{\eta},\zeta)|_{S^{\rm r}}^2&=\varpi^{\rm r}\sum_{K\in\mathcal{T}_h}\mu^{-1}h_K^2\|\nabla_h\cdot\mathbb{r}(\boldsymbol{\eta},\zeta)\|_{\boldsymbol{L}^2(K)}^2\\&\le c\sum_{K\in\mathcal{T}_h}h_K^2\left(\mu|\boldsymbol{\eta}|_{\boldsymbol{H}^2(K)}^2+\mu^{-1}|\zeta|_{\boldsymbol{H}^1(K)}^2\right)\\&\le c\,h^2\big(\mu|\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)}^2+\mu^{-1}|\zeta|_{H^1(D)}^2\big)^2.\end{aligned}$$

We have $|\zeta|_{S^{\rm p}}=0$ since $\zeta\in H^1(D)$, and $|(\boldsymbol{\eta},\zeta)|_{S^{\rm n}}=0$ because $\mathbb{r}(\boldsymbol{\eta},\zeta)_{|\partial D_{\rm n}}\boldsymbol{n}=\boldsymbol{0}$. In conclusion, we have

$$|(\boldsymbol{\eta},\zeta)|_S\le c\,h\big(\mu^{\frac12}|\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)}+\mu^{-\frac12}|\zeta|_{H^1(D)}\big).$$

(ii) We estimate $|(\boldsymbol{\eta}-\boldsymbol{\mathcal{I}}_{h\rm d}^{\rm u}(\boldsymbol{\eta}),\zeta-\mathcal{I}_h^{\rm p}(\zeta))|_{S^{\rm r}}$ by bounding the three seminorms. Concerning $|\cdot|_{S^{\rm r}}$, the definition of the stabilizing bilinear form $s_h^{\rm r}$ in (62.10a) and the approximation properties (62.24) lead to

$$\begin{aligned}|(\boldsymbol{\eta}-\boldsymbol{\mathcal{I}}_{h\rm d}^{\rm u}(\boldsymbol{\eta}),\zeta-\mathcal{I}_h^{\rm p}(\zeta))|_{S^{\rm r}}^2&\le c\sum_{K\in\mathcal{T}_h}\mu^{-1}h_K^2\|\nabla_h\cdot\mathbb{r}(\boldsymbol{\eta}-\boldsymbol{\mathcal{I}}_{h\rm d}^{\rm u}(\boldsymbol{\eta}),\zeta-\mathcal{I}_h^{\rm p}(\zeta))\|_{\boldsymbol{L}^2(K)}^2\\&\le c\sum_{K\in\mathcal{T}_h}h_K^2\left(\mu|\boldsymbol{\eta}-\boldsymbol{\mathcal{I}}_{h\rm d}^{\rm u}(\boldsymbol{\eta})|_{\boldsymbol{H}^2(K)}^2+\mu^{-1}|\zeta-\mathcal{I}_h^{\rm p}(\zeta)|_{\boldsymbol{H}^1(K)}^2\right)\\&\le c\,h^2\big(\mu|\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)}^2+\mu^{-1}|\zeta|_{H^1(D)}^2\big)^2.\end{aligned}$$

Concerning $|\cdot|_{S^{\rm p}}$, we use the triangle inequality to bound the jump norm by the norms of the traces from both sides, the definition of the stabilizing bilinear form $s_h^{\rm p}$ in (62.10b), the multiplicative trace inequality (12.16), the regularity of the mesh sequence, and the approximation property (62.24b) to infer that

$$\begin{aligned}|\zeta-\mathcal{I}_h^{\rm p}(\zeta)|_{S^{\rm p}}^2&=\sum_{F\in\mathcal{F}_h^\circ}\varpi^{\rm p}\frac{h_F}{\mu}\|[\![\zeta-\mathcal{I}_h^{\rm p}(\zeta)]\!]\|_{L^2(F)}^2\le c\,\mu^{-1}\sum_{K\in\mathcal{T}_h}h_K\|\zeta-\mathcal{I}_h^{\rm p}(\zeta)\|_{L^2(\partial K)}^2\\&\le c'\mu^{-1}\sum_{K\in\mathcal{T}_h}\|\zeta-\mathcal{I}_h^{\rm p}(\zeta)\|_{L^2(K)}(\|\zeta-\mathcal{I}_h^{\rm p}(\zeta)\|_{L^2(K)}+h_K|\zeta-\mathcal{I}_h^{\rm p}(\zeta)|_{H^1(K)})\\&\le c''\,\mu^{-1}h^2|\zeta|_{H^1(D)}^2.\end{aligned}$$

Concerning $|\cdot|_{S^{\rm n}}$, we invoke similar arguments to obtain

$$\begin{aligned}|(\boldsymbol{\eta}-\boldsymbol{\mathcal{I}}^{\rm u}_{h{\rm d}}(\boldsymbol{\eta}),\zeta-\mathcal{I}^{\rm p}_h(\zeta))|^2_{S^{\rm n}} &= \sum_{F\in\mathcal{F}^{\rm n}_h}\varpi^{\rm n}\frac{h_F}{\mu}\|\mathbb{r}(\boldsymbol{\eta}-\boldsymbol{\mathcal{I}}^{\rm u}_{h{\rm d}}(\boldsymbol{\eta}),\zeta-\mathcal{I}^{\rm p}_h(\zeta))\boldsymbol{n}\|^2_{\boldsymbol{L}^2(F)}\\ &\le c\,\mu^{-1}\sum_{K\in\mathcal{T}^{\rm n}_h}h_K\big(\mu\|\nabla(\boldsymbol{\eta}-\boldsymbol{\mathcal{I}}^{\rm u}_{h{\rm d}}(\boldsymbol{\eta}))\|_{\mathbb{L}^2(\partial K)}+\|\zeta-\mathcal{I}^{\rm p}_h(\zeta)\|_{L^2(\partial K)}\big)^2\\ &\le c'\,h^2\big(\mu|\boldsymbol{\eta}|^2_{\boldsymbol{H}^2(D)}+\mu^{-1}|\zeta|^2_{H^1(D)}\big)^2.\end{aligned}$$

Hence, we have

$$|(\boldsymbol{\eta}-\boldsymbol{\mathcal{I}}^{\rm u}_{h{\rm d}}(\boldsymbol{\eta}),\zeta-\mathcal{I}^{\rm p}_h(\zeta))|_S\le c\,h\big(\mu^{\frac12}|\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)}+\mu^{-\frac12}|\zeta|_{H^1(D)}\big).$$

(iii) By using the triangle inequality, we conclude that

$$|(\boldsymbol{\mathcal{I}}^{\rm u}_{h{\rm d}}(\boldsymbol{\eta}),\mathcal{I}^{\rm p}_h(\zeta))|_S\le c\,h\big(\mu^{\frac12}|\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)}+\mu^{-\frac12}|\zeta|_{H^1(D)}\big).$$

Exercise 62.4 (Inf-sup condition on t_h). Let $(\boldsymbol{v}_h,q_h)\in Y_h$. Let us set

$$\mathbb{S}:=\sup_{(\boldsymbol{w}_h,r_h)\in Y_h}\frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,r_h))|}{\|(\boldsymbol{w}_h,r_h)\|_{Y_h}}.$$

Using the hint and the inequality

$$\beta_0\,\mu^{-\frac12}\|q_h\|_{L^2(D)}\le\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h{\rm d}}}\frac{|b(\boldsymbol{v}_h,q_h)|}{\mu^{\frac12}|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}}+|q_h|_{S^{\rm gp}}+|q_h|_{S^{\rm p}},$$

with $\beta_0>0$, we infer that

$$\begin{aligned}\beta_0\,\mu^{-\frac12}\|q_h\|_{L^2(D)}&\le\sup_{\boldsymbol{w}_h\in\boldsymbol{V}_{h{\rm d}}}\frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0))-a(\boldsymbol{v}_h,\boldsymbol{w}_h)-s_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0))|}{\mu^{\frac12}|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}}\\ &\quad+|q_h|_{S^{\rm gp}}+|q_h|_{S^{\rm p}}\\ &\le\mathfrak{T}_1+\mathfrak{T}_2+\mathfrak{T}_3+|q_h|_{S^{\rm gp}}+|q_h|_{S^{\rm p}},\end{aligned}$$

where

$$\begin{aligned}\mathfrak{T}_1&:=\sup_{\boldsymbol{w}_h\in\boldsymbol{V}_{h{\rm d}}}\frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0))|}{\mu^{\frac12}|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}},\\ \mathfrak{T}_2&:=\sup_{\boldsymbol{w}_h\in\boldsymbol{V}_{h{\rm d}}}\frac{|a(\boldsymbol{v}_h,\boldsymbol{w}_h)|}{\mu^{\frac12}|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}},\\ \mathfrak{T}_3&:=\sup_{\boldsymbol{w}_h\in\boldsymbol{V}_{h{\rm d}}}\frac{|s_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0))|}{\mu^{\frac12}|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}}.\end{aligned}$$

Since $\partial D_{\rm d}=\partial D$ by assumption, we have $s_h=s_h^{\rm r}+s_h^{\rm p}$ (i.e., $s_h^{\rm n}:=0$), so that

$$\|(\boldsymbol{v}_h,q_h)\|^2_{Y_h}=\mu|\boldsymbol{v}_h|^2_{\boldsymbol{H}^1(D)}+\frac{1}{\mu}\|q_h\|^2_{L^2(D)}+|(\boldsymbol{v}_h,q_h)|^2_{S^{\rm r}}+|q_h|^2_{S^{\rm p}}.$$

Invoking an inverse inequality, we infer that

$$\|(\boldsymbol{w}_h,0)\|^2_{Y_h}=\mu|\boldsymbol{w}_h|^2_{\boldsymbol{H}^1(D)}+|(\boldsymbol{w}_h,0)|^2_{S^{\rm r}}\le c\,\mu|\boldsymbol{w}_h|^2_{\boldsymbol{H}^1(D)}.$$

This implies that $\mathfrak{T}_1 \le c\mathbb{S}$. Moreover, the boundedness of the bilinear form a implies that $\mathfrak{T}_2 \le 2\mu^{\frac12}|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}$. Finally, since we have $s_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0)) = s_h^{\mathrm{r}}((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0))$ and $|(\boldsymbol{w}_h,0)|_{S^{\mathrm{r}}} \le c\mu^{\frac12}|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}$, we infer that $\mathfrak{T}_3 \le c|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{r}}}$. Putting these bounds together yields

$$\begin{aligned}\mu^{-1}\|q_h\|_{L^2(D)}^2 &\le c\left(\mathbb{S}^2 + \mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2 + |(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{r}}}^2 + |q_h|_{S^{\mathrm{gp}}}^2 + |q_h|_{S^{\mathrm{p}}}^2\right)\\ &= c\left(\mathbb{S}^2 + \mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2 + |(\boldsymbol{v}_h,q_h)|_{S}^2 + |q_h|_{S^{\mathrm{gp}}}^2\right).\end{aligned}$$

Invoking a discrete trace inequality, an inverse inequality, and the regularity of the mesh sequence, we infer that (the value of c changes at each occurrence)

$$\begin{aligned}|q_h|_{S^{\mathrm{gp}}}^2 &= \sum_{F\in\mathcal{F}_h^\circ}\frac{h_F^3}{\mu}\|[\![\nabla_h q_h]\!]\cdot\boldsymbol{n}_F\|_{L^2(F)}^2 \le c\sum_{K\in\mathcal{T}_h}\frac{h_K^2}{\mu}\|\nabla_h q_h\|_{\boldsymbol{L}^2(K)}^2\\ &\le c\sum_{K\in\mathcal{T}_h}\frac{h_K^2}{\mu}\left(\|\nabla\cdot\mathbb{r}(\boldsymbol{v}_h,q_h)\|_{\boldsymbol{L}^2(K)}^2 + \mu^2\|\nabla\cdot\mathbb{e}(\boldsymbol{v}_h)\|_{\boldsymbol{L}^2(K)}^2\right)\\ &\le c\left(|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{r}}}^2 + \mu\|\mathbb{e}(\boldsymbol{v}_h)\|_{\mathbb{L}^2(K)}^2\right)\\ &\le c\left(|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{r}}}^2 + \mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2\right).\end{aligned}$$

As a result, we have

$$\mu^{-1}\|q_h\|_{L^2(D)}^2 \le c\left(\mathbb{S}^2 + \mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2 + |(\boldsymbol{v}_h,q_h)|_{S}^2\right).$$

Recalling (62.15), i.e.,

$$\alpha\left(\mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2 + |(\boldsymbol{v}_h,q_h)|_{S}^2\right) \le \mathbb{S}\|(\boldsymbol{v}_h,q_h)\|_{Y_h},$$

we obtain $\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^2 \le c(\mathbb{S}^2 + \mathbb{S}\|(\boldsymbol{v}_h,q_h)\|_{Y_h})$, and we conclude as usual by invoking Young's inequality.

Chapter 63

Stokes equations: Other stabilizations

Exercises

Exercise 63.1 (Coercivity, CIP). Prove Lemma 63.2. (*Hint*: see the proofs of Lemma 37.2 and Lemma 37.3.)

Exercise 63.2 (Inf-sup condition on b, CIP). Prove the inf-sup condition (63.13) on b. Here, we do not assume that Q_h is H^1-conforming, that is, the pressure space is either $P^{\mathrm{g}}_{k_p}(\mathcal{T}_h)$ or $P^{\mathrm{b}}_{k_p}(\mathcal{T}_h)$. (*Hint*: use the identities for $\mu^{-1}h^2\|\nabla_h q_h\|^2_{\boldsymbol{L}^2(D)}$ and $\mu^{-1}\|q_h\|^2_{L^2(D)}$ from the proof of Lemma 63.3.)

Exercise 63.3 (Galerkin orthogonality, dG). Prove the Galerkin orthogonality for the stabilized dG formulation from §63.2, i.e., $t_h((\boldsymbol{u},p),(\boldsymbol{w}_h,r_h)) = \ell_h(\boldsymbol{w}_h,r_h)$ for all $(\boldsymbol{w}_h,r_h)\in Y_h$.

Exercise 63.4 (Integration by parts for b_h, dG). Let b_h be defined in (63.19). Prove the identity (63.27). (*Hint*: $[\![ab]\!] = \{a\}[\![b]\!] + [\![a]\!]\{b\}$ at all the interfaces.)

Exercise 63.5 (dG fluxes). Derive local formulations of the discrete problem using the fluxes from Remark 63.7. (*Hint*: proceed as in §38.4.)

Exercise 63.6 (Inf-sup conditions, dG). Assume that $\partial D = \partial D_{\mathrm{d}}$. (i) Prove the inf-sup condition (63.29) on b_h. (*Hint*: use (63.26).) (ii) Using the inf-sup condition on b_h, prove again the inf-sup condition on t_h. (*Hint*: use the identity (63.28).)

Solution to exercises

Exercise 63.1 (Proof of Lemma 63.2). Let $\mathcal{T}_h^{\partial D}$ be the collection of the mesh cells having at least one boundary face, i.e., $\mathcal{T}_h^{\partial D} := \bigcup_{F\in\mathcal{F}_h^\partial}\{K_l\}$. Let us set $D^\partial := \mathrm{int}\left(\bigcup_{K\in\mathcal{T}_h^{\partial D}} K\right)$. Proceeding as in the proof of Lemma 37.2, we obtain

$$|n_h(\boldsymbol{v}_h,\boldsymbol{w}_h)| \le n_\partial^{\frac12} c_{\mathrm{dt}} (2\mu)^{\frac12} \|\mathrm{e}(\boldsymbol{v}_h)\|_{\mathbb{L}^2(D^\partial)} |\boldsymbol{w}_h|_{S^{\mathrm{u}}}.$$

This, in turn, implies that

$$a_h(\boldsymbol{v}_h,\boldsymbol{v}_h)+s_h^{\mathrm{u}}(\boldsymbol{v}_h,\boldsymbol{v}_h)\geq (x^\circ)^2+(x^\partial)^2-2n_\partial^{\frac12}c_{\mathrm{dt}}x^\partial y+\varpi^{\mathrm{u}}y^2,$$

with $x^\circ:=(2\mu)^{\frac12}\|\mathbb{e}(\boldsymbol{v}_h)\|_{\mathbb{L}^2(D\setminus D^\partial)}$, $x^\partial:=(2\mu)^{\frac12}\|\mathbb{e}(\boldsymbol{v}_h)\|_{\mathbb{L}^2(D^\partial)}$, and $y:=|\boldsymbol{v}_h|_{S^{\mathrm{u}}}$. We then infer (63.8) by proceeding as in the proof of Lemma 37.3.

Exercise 63.2 (Inf-sup condition on b, CIP). Let $q_h\in Q_h$ and set $\mathbb{B}:=\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_h}\frac{|b(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}}$. Let us set

$$\boldsymbol{w}_h:=\mu^{-1}h^2\mathcal{J}_h^{\mathrm{g,av}}(\nabla_h q_h),$$

where $\mathcal{J}_h^{\mathrm{g,av}}$ is the H^1-conforming averaging operator from §22.2. Since $\partial D_{\mathrm{d}}=\partial D$, integrating by parts gives

$$\begin{aligned}\mu^{-1}h^2\|\nabla_h q_h\|_{\boldsymbol{L}^2(D)}^2&=\mu^{-1}h^2(\nabla_h q_h-\mathcal{J}_h^{\mathrm{g,av}}(\nabla_h q_h),\nabla_h q_h)_{\boldsymbol{L}^2(D)}+b(\boldsymbol{w}_h,q_h)\\&\quad+\sum_{F\in\mathcal{F}_h^\circ}(\llbracket q_h\rrbracket\boldsymbol{n}_F,\boldsymbol{w}_h)_{\boldsymbol{L}^2(F)}=:\mathfrak{T}_1+\mathfrak{T}_2+\mathfrak{T}_3.\end{aligned}$$

Reasoning as in the proof of Lemma 63.3, and in particular using that $\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h}\leq c\mu^{-\frac12}h\|\nabla_h q_h\|_{\boldsymbol{L}^2(D)}$, we infer that

$$|\mathfrak{T}_1+\mathfrak{T}_3|\leq c(|q_h|_{S^{\mathrm{gp}}}+|q_h|_{S^{\mathrm{p}}})\mu^{-\frac12}h\|\nabla_h q_h\|_{\boldsymbol{L}^2(D)}.$$

Moreover, we have

$$|\mathfrak{T}_2|\leq\mathbb{B}\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h}\leq c\,\mathbb{B}\mu^{-\frac12}h\|\nabla_h q_h\|_{\boldsymbol{L}^2(D)}.$$

Combining these two bounds leads to

$$\mu^{-\frac12}h\|\nabla_h q_h\|_{\boldsymbol{L}^2(D)}\leq c\,(\mathbb{B}+|q_h|_{S^{\mathrm{gp}}}+|q_h|_{S^{\mathrm{p}}}).$$

Moreover, let $\boldsymbol{w}_{q_h}\in\boldsymbol{V}_{\mathrm{d}}$ be the function introduced in (62.16), i.e.,

$$\nabla\cdot\boldsymbol{w}_{q_h}=-\mu^{-1}q_h,\qquad\beta_D|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)}\leq\mu^{-1}\|q_h\|_{L^2(D)},$$

with $\beta_D>0$. Letting $\boldsymbol{w}_h:=\boldsymbol{\mathcal{I}}_h^{\mathrm{g,av}}(\boldsymbol{w}_{q_h})\in\boldsymbol{V}_h$, we have

$$\begin{aligned}\mu^{-1}\|q_h\|_{L^2(D)}^2&=(\nabla_h q_h,\boldsymbol{w}_{q_h}-\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)}+b(\boldsymbol{w}_h,q_h)\\&\quad+\sum_{F\in\mathcal{F}_h^\circ}(\llbracket q_h\rrbracket\boldsymbol{n}_F,\boldsymbol{w}_h-\boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(F)}=:\mathfrak{T}_1+\mathfrak{T}_2+\mathfrak{T}_3.\end{aligned}$$

Reasoning as in the proof of Lemma 63.3 and using the above bound on $\mu^{-\frac12}h\|\nabla_h q_h\|_{\boldsymbol{L}^2(D)}$ readily yields

$$\mu^{-\frac12}\|q_h\|_{L^2(D)}\leq c\,(\mathbb{B}+|q_h|_{S^{\mathrm{gp}}}+|q_h|_{S^{\mathrm{p}}}),$$

which is the expected estimate on $\mu^{-\frac12}\|q_h\|_{L^2(D)}$.

Exercise 63.3 (Galerkin orthogonality, dG). Using the boundary condition $\mathbb{r}(\boldsymbol{u},p)_{|F}\boldsymbol{n}=\boldsymbol{0}$ for all $F\in\mathcal{F}_h^{\mathrm{n}}$, and using that $\mathbb{r}(\boldsymbol{u},p)\boldsymbol{n}$ is continuous across the mesh interfaces, we have

$$\begin{aligned}\ell_h(\boldsymbol{w}_h,r_h)=\ell(\boldsymbol{w}_h,r_h)&=(\nabla\cdot\mathbb{r}(\boldsymbol{u},p),\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)}+(\nabla\cdot\boldsymbol{u},r_h)_{L^2(D)}\\&=(\mathbb{s}(\boldsymbol{u}),\mathbb{e}_h(\boldsymbol{w}_h))_{\boldsymbol{L}^2(D)}-(\nabla_h\cdot\boldsymbol{w}_h,p)_{L^2(D)}+(\nabla\cdot\boldsymbol{u},r_h)_{L^2(D)}\\&\quad+\sum_{F\in\mathcal{F}_h^\circ\cup\mathcal{F}_h^{\mathrm{d}}}(\{\mathbb{r}(\boldsymbol{u},p)\}\boldsymbol{n}_F,\llbracket\boldsymbol{w}_h\rrbracket)_{\boldsymbol{L}^2(F)}.\end{aligned}$$

Using that $[\![\boldsymbol{u}]\!]_F = \boldsymbol{0}$ for all $F \in \mathcal{F}_h^\circ \cup \mathcal{F}_h^{\mathrm{d}}$, $s_h^{\mathrm{p}}(p, q_h) = 0$ for all $q_h \in Q_h$, and $s_h^{\mathrm{n}}((\boldsymbol{u}, p), (\boldsymbol{w}_h, r_h)) = 0$ for all $(\boldsymbol{w}_h, r_h) \in Y_h$, we infer that

$$\begin{aligned}
\ell_h(\boldsymbol{w}_h, r_h) &= (\mathbb{s}(\boldsymbol{u}), \mathbb{e}_h(\boldsymbol{w}_h))_{\boldsymbol{L}^2(D)} - \sum_{F \in \mathcal{F}_h^\circ \cup \mathcal{F}_h^{\mathrm{d}}} (\{\mathbb{s}(\boldsymbol{u})\}\boldsymbol{n}_F, [\![\boldsymbol{w}_h]\!])_{\boldsymbol{L}^2(F)} \\
&\quad - \sum_{F \in \mathcal{F}_h^\circ \cup \mathcal{F}_h^{\mathrm{d}}} (\{\mathbb{s}(\boldsymbol{w}_h)\}\boldsymbol{n}_F, [\![\boldsymbol{u}]\!])_{\boldsymbol{L}^2(F)} + s_h^{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{w}_h) \\
&\quad - (\nabla_h \cdot \boldsymbol{w}_h, p)_{L^2(D)} + \sum_{F \in \mathcal{F}_h^\circ \cup \mathcal{F}_h^{\mathrm{d}}} ([\![\boldsymbol{w}_h]\!] \cdot \boldsymbol{n}_F, \{p\})_{L^2(F)} \\
&\quad + (\nabla_h \cdot \boldsymbol{u}, r_h)_{L^2(D)} - \sum_{F \in \mathcal{F}_h^\circ \cup \mathcal{F}_h^{\mathrm{d}}} ([\![\boldsymbol{u}]\!] \cdot \boldsymbol{n}_F, \{r_h\})_{\boldsymbol{L}^2(F)} \\
&= a_h(\boldsymbol{u}, \boldsymbol{w}_h + s_h^{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{w}_h) + b_h(\boldsymbol{w}_h, p) - b_h(\boldsymbol{u}, r_h) \\
&= t_h((\boldsymbol{u}, p), (\boldsymbol{w}_h, r_h)).
\end{aligned}$$

We have thus proved that $t_h((\boldsymbol{u}, p), (\boldsymbol{w}_h, r_h)) = \ell_h(\boldsymbol{w}_h, r_h)$.

Exercise 63.4 (Integration by parts for b_h, dG). Integrating by parts elementwise, we have

$$\begin{aligned}
b_h(\boldsymbol{v}_h, q_h) &= -(\nabla_h \cdot \boldsymbol{v}_h, q_h)_{L^2(D)} + \sum_{F \in \mathcal{F}_h^\circ \cup \mathcal{F}_h^{\mathrm{d}}} ([\![\boldsymbol{v}_h]\!] \cdot \boldsymbol{n}_F, \{q_h\})_{L^2(F)} \\
&= (\boldsymbol{v}_h, \nabla_h q_h)_{\boldsymbol{L}^2(D)} - \sum_{K \in \mathcal{T}_h} (\boldsymbol{v}_{h|K} \cdot \boldsymbol{n}_K, q_{h|K})_{L^2(\partial K)} \\
&\quad + \sum_{F \in \mathcal{F}_h^\circ \cup \mathcal{F}_h^{\mathrm{d}}} ([\![\boldsymbol{v}_h]\!] \cdot \boldsymbol{n}_F, \{q_h\})_{L^2(F)},
\end{aligned}$$

recalling that the jump and average operators return the actual value at boundary faces. We observe that

$$\begin{aligned}
\sum_{K \in \mathcal{T}_h} (\boldsymbol{v}_{h|K} \cdot \boldsymbol{n}_K, q_{h|K})_{L^2(\partial K)} &= \sum_{F \in \mathcal{F}_h^\circ} ([\![\boldsymbol{v}_h q_h]\!] \cdot \boldsymbol{n}_F, 1)_{L^2(F)} + \sum_{F \in \mathcal{F}_h^{\mathrm{d}} \cup \mathcal{F}_h^{\mathrm{n}}} (\boldsymbol{v}_h \cdot \boldsymbol{n}, q_h)_{L^2(F)} \\
&= \sum_{F \in \mathcal{F}_h^\circ} ([\![\boldsymbol{v}_h]\!] \cdot \boldsymbol{n}_F, \{q_h\})_{L^2(F)} + \sum_{F \in \mathcal{F}_h^\circ} (\{\boldsymbol{v}_h\} \cdot \boldsymbol{n}_F, [\![q_h]\!])_{L^2(F)} \\
&\quad + \sum_{F \in \mathcal{F}_h^{\mathrm{d}} \cup \mathcal{F}_h^{\mathrm{n}}} (\boldsymbol{v}_h \cdot \boldsymbol{n}, q_h)_{L^2(F)},
\end{aligned}$$

where the second equality follows by using the hint. Combining the above two identities gives

$$\begin{aligned}
b_h(\boldsymbol{v}_h, q_h) &= (\boldsymbol{v}_h, \nabla_h q_h)_{\boldsymbol{L}^2(D)} - \sum_{F \in \mathcal{F}_h^\circ} (\{\boldsymbol{v}_h\} \cdot \boldsymbol{n}_F, [\![q_h]\!])_{L^2(F)} - \sum_{F \in \mathcal{F}_h^{\mathrm{n}}} (\boldsymbol{v}_h \cdot \boldsymbol{n}, q_h)_{L^2(F)} \\
&= (\boldsymbol{v}_h, \nabla_h q_h)_{\boldsymbol{L}^2(D)} - \sum_{F \in \mathcal{F}_h^\circ \cup \mathcal{F}_h^{\mathrm{n}}} (\{\boldsymbol{v}_h\} \cdot \boldsymbol{n}_F, [\![q_h]\!])_{L^2(F)},
\end{aligned}$$

using again the above convention on the jump and average operators associated with the boundary faces. This proves the expected identity.

Exercise 63.5 (dG fluxes). Let $K \in \mathcal{T}_h$, let $\mathbb{1}_K$ be the indicator function of K, and let $\boldsymbol{\xi} \in \mathbb{P}_{k_u}$. Let us use $(\boldsymbol{\xi}\mathbb{1}_K, 0)$ as a test function in the discrete problem. We obtain

$$(\boldsymbol{f}, \boldsymbol{\xi})_{\boldsymbol{L}^2(K)} = a_h(\boldsymbol{u}_h, \boldsymbol{\xi}\mathbb{1}_K) + b_h(\boldsymbol{\xi}\mathbb{1}_K, p_h) + s_h^{\mathrm{u}}(\boldsymbol{u}_h, \boldsymbol{\xi}\mathbb{1}_K) + s_h^{\mathrm{n}}((\boldsymbol{u}_h, p_h), (\boldsymbol{\xi}\mathbb{1}_K, 0)).$$

Since $[\![\boldsymbol{\xi}\mathbb{1}_K]\!]_F = \epsilon_{K,F}\boldsymbol{\xi}$ and $(\mathbb{L}_F^l(\boldsymbol{v}), \mathfrak{q}_h)_{\mathbb{L}^2(D)} = (\boldsymbol{v}, \{\mathfrak{q}_h\}\boldsymbol{n}_F)_{\boldsymbol{L}^2(F)}$, for all $\boldsymbol{v} \in \boldsymbol{L}^2(F)$, all $\mathfrak{q}_h \in P_l^{\rm b}(\mathcal{T}_h;\mathbb{R}^{d\times d})$, and all $F \in \mathcal{F}_h$, we have

$$\begin{aligned} a_h(\boldsymbol{u}_h, \boldsymbol{\xi}\mathbb{1}_K) &= (\mathbb{s}_h(\boldsymbol{u}_h), \mathbb{e}(\boldsymbol{\xi}))_{\mathbb{L}^2(K)} - n_h(\boldsymbol{u}_h, \boldsymbol{\xi}\mathbb{1}_K) - n_h(\boldsymbol{\xi}\mathbb{1}_K, \boldsymbol{u}_h) \\ &= (\mathbb{s}_h(\boldsymbol{u}_h), \mathbb{e}(\boldsymbol{\xi}))_{\mathbb{L}^2(K)} - \sum_{F\in\mathcal{F}_K\cap(\mathcal{F}_h^\circ\cup\mathcal{F}_h^{\rm d})} (\{\mathbb{s}_h(\boldsymbol{u}_h)\}\boldsymbol{n}_F, \epsilon_{K,F}\boldsymbol{\xi})_{\boldsymbol{L}^2(F)} \\ &\quad - \sum_{F\in\mathcal{F}_K\cap(\mathcal{F}_h^\circ\cup\mathcal{F}_h^{\rm d})} (\mathbb{L}_F^l([\![\boldsymbol{u}_h]\!]), 2\mu\mathbb{e}(\boldsymbol{\xi}))_{\mathbb{L}^2(K)}, \end{aligned}$$

$$\begin{aligned} b_h(\boldsymbol{\xi}\mathbb{1}_K, p_h) &= -(p_h, \nabla\cdot\boldsymbol{\xi})_{L^2(K)} + \sum_{F\in\mathcal{F}_K\cap(\mathcal{F}_h^\circ\cup\mathcal{F}_h^{\rm d})} (\{p_h\}, \epsilon_{K,F}\boldsymbol{\xi}\cdot\boldsymbol{n}_F)_{L^2(F)} \\ &= -(p_h\mathbb{I}, \mathbb{e}(\boldsymbol{\xi}))_{\mathbb{L}^2(K)} + \sum_{F\in\mathcal{F}_K\cap(\mathcal{F}_h^\circ\cup\mathcal{F}_h^{\rm d})} (\{p_h\mathbb{I}\}\boldsymbol{n}_F, \epsilon_{K,F}\boldsymbol{\xi})_{\boldsymbol{L}^2(F)}, \end{aligned}$$

$$s_h^{\rm u}(\boldsymbol{u}_h, \boldsymbol{\xi}\mathbb{1}_K) = \sum_{F\in\mathcal{F}_K\cap(\mathcal{F}_h^\circ\cup\mathcal{F}_h^{\rm d})} \varpi^{\rm u}\frac{2\mu}{h_F}([\![\boldsymbol{u}_h]\!], \epsilon_{K,F}\boldsymbol{\xi})_{\boldsymbol{L}^2(F)},$$

and

$$\begin{aligned} s_h^{\rm n}((\boldsymbol{u}_h, p_h), (\boldsymbol{\xi}\mathbb{1}_K, 0)) &= \sum_{F\in\mathcal{F}_K\cap\mathcal{F}_h^{\rm n}} \varpi^{\rm n}\frac{h_F}{\mu}(\mathbb{r}_h(\boldsymbol{u}_h, p_h)\boldsymbol{n}, -(2\mu)\mathbb{e}(\boldsymbol{\xi})\boldsymbol{n})_{\boldsymbol{L}^2(F)} \\ &= -\sum_{F\in\mathcal{F}_K\cap\mathcal{F}_h^{\rm n}} 2\varpi^{\rm n}h_F(\mathbb{L}_F^l(\mathbb{r}_h(\boldsymbol{u}_h, p_h)\boldsymbol{n}), \mathbb{e}(\boldsymbol{\xi}))_{\mathbb{L}^2(K)}. \end{aligned}$$

Recalling the definition of the global lifting

$$\mathbb{L}_h^l(\boldsymbol{u}_h, p_h) := \sum_{F\in\mathcal{F}_h^\circ\cup\mathcal{F}_h^{\rm d}} \mathbb{L}_F^l([\![\boldsymbol{u}_h]\!]) + \sum_{F\in\mathcal{F}_h^{\rm n}} \varpi^{\rm n}\frac{h_F}{\mu}\mathbb{L}_F^l(\mathbb{r}_h(\boldsymbol{u}_h, p_h)\boldsymbol{n}),$$

and the definition of the discrete total stress tensor $\widetilde{\mathbb{r}}_h^l(\boldsymbol{u}_h, p_h) := \mathbb{r}_h(\boldsymbol{u}_h, p_h) + 2\mu\mathbb{L}_h^l(\boldsymbol{u}_h, p_h)$, this leads to

$$-(\widetilde{\mathbb{r}}_h^l(\boldsymbol{u}_h, p_h), \mathbb{e}(\boldsymbol{\xi}))_{\mathbb{L}^2(K)} + \sum_{F\in\mathcal{F}_K} \epsilon_{K,F}(\boldsymbol{\Phi}_F^{\rm u}(\boldsymbol{u}_h, p_h), \boldsymbol{\xi})_{\boldsymbol{L}^2(F)} = (\boldsymbol{f}, \boldsymbol{\xi})_{\boldsymbol{L}^2(K)},$$

with the flux

$$\boldsymbol{\Phi}_F^{\rm u}(\boldsymbol{u}_h, p_h) := \begin{cases} \{\mathbb{r}_h(\boldsymbol{u}_h, p_h)\}\boldsymbol{n}_F + \varpi^{\rm u}\frac{2\mu}{h_F}[\![\boldsymbol{u}_h]\!] & \text{if } F \in \mathcal{F}_h^\circ\cup\mathcal{F}_h^{\rm d}, \\ \boldsymbol{0} & \text{if } F \in \mathcal{F}_h^{\rm n}. \end{cases}$$

Let now $\zeta \in \mathbb{P}_{k_p}$ and let us use $(\boldsymbol{0}, \zeta\mathbb{1}_K)$ as a test function in the discrete problem. We obtain

$$(g, \zeta)_{L^2(K)} = -b_h(\boldsymbol{u}_h, \zeta\mathbb{1}_K) + s_h^{\rm p}(p_h, \zeta\mathbb{1}_K) + s_h^{\rm n}((\boldsymbol{u}_h, p_h), (\boldsymbol{0}, \zeta\mathbb{1}_K)).$$

Using the identity (63.27), we observe that

$$-b_h(\boldsymbol{u}_h, \zeta\mathbb{1}_K) = -(\boldsymbol{u}_h, \nabla\zeta)_{\boldsymbol{L}^2(K)} + \sum_{F\in\mathcal{F}_K\cap(\mathcal{F}_h^\circ\cup\mathcal{F}_h^{\rm n})} (\{\boldsymbol{u}_h\}\cdot\boldsymbol{n}_F, \epsilon_{K,F}\zeta)_{L^2(F)},$$

$$s_h^{\mathrm{p}}(p_h,\zeta\mathbb{1}_K)=\sum_{F\in\mathcal{F}_K\cap\mathcal{F}_h^\circ}\varpi^{\mathrm{p}}\frac{h_F}{\mu}(\llbracket p_h\rrbracket,\epsilon_{K,F}\zeta)_{L^2(F)},$$

and

$$s_h^{\mathrm{n}}((\boldsymbol{u}_h,p_h),(\boldsymbol{0},\zeta\mathbb{1}_K))=\sum_{F\in\mathcal{F}_K\cap\mathcal{F}_h^{\mathrm{n}}}\varpi^{\mathrm{n}}\frac{h_F}{\mu}(\mathbb{r}_h(\boldsymbol{u}_h,p_h)\boldsymbol{n},\zeta\boldsymbol{n})_{\boldsymbol{L}^2(F)}.$$

Altogether, we obtain

$$-(\boldsymbol{u}_h,\nabla\zeta)_{\boldsymbol{L}^2(K)}+\sum_{F\in\mathcal{F}_K}\epsilon_{K,F}(\Phi_F^{\mathrm{p}}(\boldsymbol{u}_h,p_h),\zeta)_{L^2(F)}=(g,\zeta)_{L^2(K)},$$

with the flux

$$\Phi_F^{\mathrm{p}}(\boldsymbol{u}_h,p_h):=\begin{cases}\{\boldsymbol{u}_h\}\cdot\boldsymbol{n}_F+\varpi^{\mathrm{p}}\frac{h_F}{\mu}\llbracket p_h\rrbracket & \text{if } F\in\mathcal{F}_h^\circ,\\ 0 & \text{if } F\in\mathcal{F}_h^{\mathrm{d}},\\ \boldsymbol{u}_h\cdot\boldsymbol{n}+\varpi^{\mathrm{n}}\frac{h_F}{\mu}\boldsymbol{n}^{\mathsf{T}}\mathbb{r}_h(\boldsymbol{u}_h,p_h)\boldsymbol{n} & \text{if } F\in\mathcal{F}_h^{\mathrm{n}}.\end{cases}$$

Exercise 63.6 (Inf-sup conditions, dG). (i) Let $q_h\in Q_h\backslash\{0\}$. Let $\boldsymbol{w}_{q_h}\in\boldsymbol{V}_{\mathrm{d}}$ be the function introduced in (62.16), i.e.,

$$\nabla\cdot\boldsymbol{w}_{q_h}=-\mu^{-1}q_h,\qquad \beta_D|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)}\le\mu^{-1}\|q_h\|_{L^2(D)},$$

with $\beta_D>0$. Let $\boldsymbol{w}_h:=\boldsymbol{\mathcal{I}}_h^{\mathrm{b}}(\boldsymbol{w}_{q_h})$ be the L^2-orthogonal projection of $\boldsymbol{w}_{q_h}$ onto $\boldsymbol{V}_h$. Recall that

$$\|\boldsymbol{v}-\boldsymbol{\mathcal{I}}_h^{\mathrm{b}}(\boldsymbol{v})\|_{\boldsymbol{L}^2(K)}+h_K|\boldsymbol{\mathcal{I}}_h^{\mathrm{b}}(\boldsymbol{v})|_{\boldsymbol{H}^1(K)}\le ch_K|\boldsymbol{v}|_{\boldsymbol{H}^1(K)},$$

for all $K\in\mathcal{T}_h$ and all $\boldsymbol{v}\in\boldsymbol{H}^1(K)$. Let us set $\mathbb{B}:=\sup_{\boldsymbol{v}_h\in\boldsymbol{V}_h}\frac{|b_h(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}}$. Recall that we have shown in (63.26) that

$$\mu^{-1}\|q_h\|_{L^2(D)}^2=b_h(\boldsymbol{w}_h,q_h)+\sum_{F\in\mathcal{F}_h^\circ}(\llbracket q_h\rrbracket\boldsymbol{n}_F,\{\boldsymbol{w}_h-\boldsymbol{w}_{q_h}\})_{\boldsymbol{L}^2(F)}=:\mathfrak{T}_1+\mathfrak{T}_2,$$

where we used that $\partial D_{\mathrm{n}}=\emptyset$ to drop the subset $\mathcal{F}_h^{\mathrm{n}}$ in $\mathfrak{T}_2$. We also used that $k_u\ge k_p$ to establish that $(\nabla_h q_h,\boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(D)}=(\nabla_h q_h,\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)}$. Owing to (63.25), we infer that

$$|\mathfrak{T}_1|\le\mathbb{B}\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h}\le c\,\mathbb{B}\mu^{-\frac12}\|q_h\|_{L^2(D)}.$$

Using the Cauchy–Schwarz inequality and since $h_F^{-\frac12}\|\boldsymbol{w}_h-\boldsymbol{w}_{q_h}\|_{\boldsymbol{L}^2(F)}\le c|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(K)}$ for all $K\in\mathcal{T}_h$ and all $F\in\mathcal{F}_K$, we have

$$|\mathfrak{T}_2|\le c\,|q_h|_{S^{\mathrm{p}}}\mu^{\frac12}|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)}\le c'\,|q_h|_{S^{\mathrm{p}}}\mu^{-\frac12}\|q_h\|_{L^2(D)}.$$

Putting everything together leads to the expected inf-sup condition (63.29) on b_h.
(ii) Let $(\boldsymbol{v}_h,q_h)\in Y_h$ and set $\mathbb{S}:=\sup_{(\boldsymbol{w}_h,r_h)\in Y_h}\frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,r_h))|}{\|(\boldsymbol{w}_h,r_h)\|_{Y_h}}$. Recall that we have

$$\alpha\left(\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}^2+|q_h|_{S^{\mathrm{p}}}^2\right)\le\mathbb{S}\|(\boldsymbol{v}_h,q_h)\|_{Y_h},$$

with $\alpha > 0$ and where we dropped the contribution from s_h^{n} since $\partial D_{\mathrm{n}} = \emptyset$. Using the hint (i.e., the identity (63.28)) and the inf-sup condition on b_h yields

$$\begin{aligned} c\,\mu^{-\frac{1}{2}}\|q_h\|_{L^2(D)} &\le \sup_{\boldsymbol{w}_h \in \boldsymbol{V}_h} \frac{|b_h(\boldsymbol{w}_h, q_h)|}{\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h}} + |q_h|_{S^{\mathrm{p}}} \\ &= \sup_{\boldsymbol{w}_h \in \boldsymbol{V}_h} \frac{|t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, 0)) - a_h(\boldsymbol{v}_h, \boldsymbol{w}_h) - s_h^{\mathrm{u}}(\boldsymbol{v}_h, \boldsymbol{w}_h)|}{\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h}} + |q_h|_{S^{\mathrm{p}}} \\ &\le \sup_{\boldsymbol{w}_h \in \boldsymbol{V}_h} \frac{|a_h(\boldsymbol{v}_h, \boldsymbol{w}_h) + s_h^{\mathrm{u}}(\boldsymbol{v}_h, \boldsymbol{w}_h)|}{\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h}} + \mathbb{S} + |q_h|_{S^{\mathrm{p}}} \le c\,\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h} + \mathbb{S} + |q_h|_{S^{\mathrm{p}}}, \end{aligned}$$

where we used the boundedness of $a_h + s_h^{\mathrm{u}}$ in the second line (which follows by using the Cauchy–Schwarz inequality and a discrete trace inequality to bound n_h) to infer that

$$\sup_{\boldsymbol{w}_h \in \boldsymbol{V}_h} \frac{|a_h(\boldsymbol{v}_h, \boldsymbol{w}_h) + s_h^{\mathrm{u}}(\boldsymbol{v}_h, \boldsymbol{w}_h)|}{\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h}} \le c\,\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h},$$

and the fact that $\|(\boldsymbol{w}_h, 0)\|_{Y_h} = \|\boldsymbol{w}_h\|_{\boldsymbol{V}_h}$ (since $\partial D_{\mathrm{n}} = \emptyset$) to infer that

$$\sup_{\boldsymbol{w}_h \in \boldsymbol{V}_h} \frac{|t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, 0))|}{\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h}} \le \mathbb{S}.$$

Putting everything together leads to $\|(\boldsymbol{v}_h, q_h)\|_{Y_h}^2 \le c(\mathbb{S}^2 + \mathbb{S}\|(\boldsymbol{v}_h, q_h)\|_{Y_h})$, and we conclude by invoking Young's inequality.

Chapter 64

Bochner integration

Exercises

Exercise 64.1 (Strong measurability). Prove the statement made in Example 64.8. (*Hint*: use Theorem 1.17.)

Exercise 64.2 (Bochner integral). Let $f : J \to V$ be a Bochner integrable function and let $(f_n)_{n\in\mathbb{N}}$ be a countable sequence of simple functions satisfying the assumptions of Definition 64.11. (i) Show that $\int_J f_n(t)\mathrm{d}t$ has a limit when $n \to \infty$. (*Hint*: prove that it is a Cauchy sequence.) (ii) Show that if $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ are two sequences of simple functions satisfying the assumptions of Definition 64.11, then $\lim_{n\to\infty} \int_J f_n(t)\mathrm{d}t = \lim_{n\to\infty} \int_J g_n(t)\mathrm{d}t$.

Exercise 64.3 ($L^p(J;V)$). Let f be a Bochner integrable function. (i) Prove that $\| \int_J f(t)\mathrm{d}t\|_V \leq \int_J \|f(t)\|_V \mathrm{d}t$. (ii) Prove that $L^p(J;V) \hookrightarrow L^1(J;V)$. (iii) Let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $L^1(J;V)$ s.t. $(f_n(t))_{n\in\mathbb{N}}$ converges to $f(t)$ in V and $\|f_n(t)\|_V \leq g(t)$ with $g \in L^1(J;\mathbb{R})$ for a.e. $t \in J$. Show that $f \in L^1(J;V)$ and $(f_n)_{n\in\mathbb{N}}$ converges to f in $L^1(J;V)$.

Exercise 64.4 ($L^q((0,1);L^p(0,1))$). Let $p \in [1,\infty)$. Let $J := (0,1)$ and $g : J \to L^p(D)$ with $D := (0,1)$ be defined by $g(t) := \mathbb{1}_{(0,t)}$ for all $t \in J$. (i) Show that g is almost separably valued. (ii) Show that g is weakly measurable. (iii) Let $q \in [1,\infty]$. Show that $g \in L^q(J;V)$ and compute $\|g\|_{L^q(J;V)}$.

Exercise 64.5 (Constants). Let V be a Banach space and $f \in L^1_{\mathrm{loc}}(J;V)$. Assume that f is weakly differentiable and $\partial_t f = 0$. Show that there is $a \in V$ such that $f(t) = a$ a.e. $t \in J$. (*Hint*: see the proof of Lemma 2.11.)

Exercise 64.6 (Linear map). Prove Lemma 64.34.

Exercise 64.7 ($X^{p,q}(J;V,W)$). Prove that $X^{p,q}(J;V,W)$ is a Banach space.

Exercise 64.8 (Continuous embedding). Let $J := (0,T)$, $T > 0$. The goal is to prove that $X^{p,q}(J;V,W) \hookrightarrow C^0(\overline{J};W)$. Let $u \in X^{p,q}(J;V,W)$. Set $v(t) := \partial_t u(t)$ and $w(t) := \int_0^t v(\tau)\,\mathrm{d}\tau$. (i) Show that $w \in C^0(\overline{J};W)$. (*Hint*: use Lebesgue's dominated convergence theorem.) (ii) Let $\rho(\tau) := \eta e^{-\frac{1}{1-|\tau|^2}}$ if $|\tau| \leq 1$ and $\rho(\tau) := 0$ otherwise, with η s.t. $\int_{\mathbb{R}} \rho(\tau)\,\mathrm{d}\tau = 1$. Let $0 < s < t < T$ and let N be the smallest integer s.t. $N \geq \max(\frac{1}{s}, \frac{1}{T-t})$. Define $\rho_n(\tau) := n\rho(n\tau)$ for all $n \geq N$. Consider

the sequence of smooth functions $\phi_n(\tau) := \int_0^\tau (\rho_n(s-\xi) - \rho_n(t-\xi))\,\mathrm{d}\xi$. What is $\lim_{n\to\infty} \phi_n(\tau)$? (*Hint*: $\int_{\mathbb{R}} \rho_n(s-\xi) f(\xi)\,\mathrm{d}\xi \to f(s)$ for a.e. s and all $f \in L^1(\mathbb{R})$.) (iii) Show that $\delta_n(s,t) := \int_{-1}^1 \rho_n(y)(u(s-\frac{y}{n}) - u(t-\frac{y}{n}))\,\mathrm{d}y = -\int_0^T v(\tau)\phi_n(\tau)\,\mathrm{d}\tau$. (iv) Compute $\lim_{n\to\infty} \delta_n(s,t)$. (*Hint*: pass to the limit in the above equality and accept as a fact that $\lim_{n\to\infty} \int_{-1}^1 \rho(\tau) f(s-\frac{\tau}{n})\,\mathrm{d}\tau = f(s)$ for a.e. s and all $f \in L^1(J;B)$, where B is either V or W.) (v) Prove that $u \in C^0(\overline{J};W)$ and $u \in C^{0,\frac{q-1}{q}}(\overline{J};W)$ if $q > 1$.

Exercise 64.9 (Time derivative of product). Let $\alpha \in C^1(\overline{J};\mathbb{R})$ and $u \in X^{p,q}(J;V,W)$. Show that $\partial_t(\alpha u) = u\partial_t\alpha + \alpha\partial_t u$ (see Definition 64.35).

Solution to exercises

Exercise 64.1 (Strong measurability). Let $i \in \{1\!:\!I\}$. Since ψ_i is integrable and reasoning on the positive and negative parts of ψ_i, we infer from Theorem 1.17 that there exists a sequence of scalar-valued simple functions $(g_{i,n})_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} g_{i,n}(t) = \psi_i(t)$ for a.e. $t \in J$. After observing that the sum of two simple functions is still a simple function, we conclude that $f_n(t) := \sum_{i\in\{1:I\}} g_{i,n}(t)\varphi_i$ is a V-valued simple function. Therefore, we have for a.e. $t \in J$,

$$\begin{aligned}\lim_{n\to\infty} \|f(t) - f_n(t)\|_V &\le \lim_{n\to\infty} \sum_{i\in\{1:I\}} |\psi_i(t) - g_{i,n}(t)|\|\varphi_i\|_V \\ &\le \sum_{i\in\{1:I\}} \lim_{n\to\infty} |\psi_i(t) - g_{i,n}(t)|\|\varphi_i\|_V = 0,\end{aligned}$$

showing that f is strongly measurable.

Exercise 64.2 (Bochner integral). (i) Let $z_n := \int_J f_n(t)\mathrm{d}t$. Lemma 64.2 implies that

$$\begin{aligned}\|z_m - z_n\|_V &\le \int_J \|f_m(t) - f_n(t)\|_V \mathrm{d}t \\ &\le \int_J (\|f_m(t) - f(t)\|_V + \|f_n(t) - f(t)\|_V)\,\mathrm{d}t \\ &= \int_J \|f_m(t) - f(t)\|_V \mathrm{d}t + \int_J \|f_n(t) - f(t)\|_V \mathrm{d}t.\end{aligned}$$

For all $\epsilon > 0$, there is $N(\epsilon)$ such that $\int_J \|f_m(t) - f(t)\|_V \mathrm{d}t + \int_J \|f_n(t) - f(t)\|_V \mathrm{d}t \le \epsilon$ for all $m, n \ge N(\epsilon)$ by assumption. This proves that $(z_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Hence, there exists $z \in V$ s.t. $z_n \to z$ as $n \to \infty$ since V is complete.
(ii) Let $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ be two sequences of simple functions satisfying the assumptions of Definition 64.11. Lemma 64.2 implies that

$$\begin{aligned}\left\|\int_J f_n(t)\mathrm{d}t - \int_J g_n(t)\mathrm{d}t\right\|_V &\le \int_J \|f_n(t) - g_n(t)\|_V \mathrm{d}t \\ &\le \int_J \|f_n(t) - f(t)\|_V \mathrm{d}t + \int_J \|g_n(t) - f(t)\|_V \mathrm{d}t.\end{aligned}$$

The assumptions of Definition 64.11 imply that

$$0 \le \limsup_{n\to\infty} \left\|\int_J f_n(t)\mathrm{d}t - \int_J g_n(t)\mathrm{d}t\right\|_V \le 0,$$

which proves the statement.

Exercise 64.3 ($L^p(J;V)$). (i) Let $(f_n)_{n\in\mathbb{N}}$ be a countable sequence of simple functions converging to f. Invoking the triangle inequality, Lemma 64.2, and again the triangle inequality, we infer that

$$\begin{aligned}\left\|\int_J f(t)\mathrm{d}t\right\|_V &\le \left\|\int_J (f(t)-f_n(t))\mathrm{d}t\right\|_V + \left\|\int_J f_n(t)\mathrm{d}t\right\|_V\\ &\le \left\|\int_J (f(t)-f_n(t))\mathrm{d}t\right\|_V + \int_J \|f_n(t)\|_V\mathrm{d}t\\ &\le \left\|\int_J (f(t)-f_n(t))\mathrm{d}t\right\|_V + \int_J \|f_n(t)-f(t)\|_V\mathrm{d}t + \int_J \|f(t)\|_V\mathrm{d}t.\end{aligned}$$

The conclusion follows from

$$\lim_{n\to\infty}\left\|\int_J (f(t)-f_n(t))\mathrm{d}t\right\|_V = 0, \qquad \lim_{n\to\infty}\int_J \|f(t)-f_n(t)\|_V\mathrm{d}t = 0,$$

which are consequences of the definition of f being Bochner integrable.
(ii) Since the function $\phi : J \ni t \longmapsto \|f(t)\|_V \in \mathbb{R}$ is integrable and J is bounded, we have

$$\|f\|_{L^1(J;V)} := \|\phi\|_{L^1(J)} \le |J|^{\frac{p-1}{p}}\|\phi\|_{L^p(J)} =: |J|^{\frac{p-1}{p}}\|f\|_{L^p(J;V)},$$

which proves the statement for $p \in [1,\infty)$. The case $p=\infty$ follows from the inequality

$$\|f\|_{L^1(J;V)} := \|\phi\|_{L^1(J)} \le \|\phi\|_{L^\infty(J)} =: \|f\|_{L^\infty(J;V)}.$$

(iii) We first show that the function f is strongly measurable. For all $k\in\mathbb{N}$, there is $n_k\in\mathbb{N}$ s.t. $\|f(t)-f_{n_k}(t)\|_V \le (k+1)^{-1}$, and since f_{n_k} is strongly measurable, there is a simple function of the form $\sum_{l\in\{1:m_k\}} v_l \mathbb{1}_{A_l}(t)$ s.t.

$$\left\|f_{h_k}(t) - \sum_{l\in\{1:m_k\}} v_l\mathbb{1}_{A_l}(t)\right\|_V \le (k+1)^{-1}.$$

Then the sequence of simple functions $(\sum_{l\in\{1:m_k\}} v_l\mathbb{1}_{A_l}(t))_{k\in\mathbb{N}}$ converges simply to f. Applying Lebesgue's dominated convergence theorem in $L^1(J;\mathbb{R})$ to the sequence of functions $(\|f_n\|_V)_{n\in\mathbb{N}}$, i.e., $\|f_n(t)\|_V \to \|f(t)\|_V$ and $\|f_n(t)\|_V \le g(t)$ for a.e. $t\in J$, we infer that $\|f\|_V$ is in $L^1(J;\mathbb{R})$. Hence, f is Bochner integrable and $f\in L^1(J;V)$. Applying again Lebesgue's dominated convergence theorem to the sequence $(\|f_n-f\|_V)_{n\in\mathbb{N}}$, i.e., $\|f_n(t)-f(t)\|_V\to 0$ and $\|f_n(t)-f(t)\|_V \le g(t)+\|f(t)\|_V$ for a.e. $t\in J$, we infer that $\int_J \|f_n(t)-f(t)\|_V\mathrm{d}t\to 0$ as $n\to\infty$. This proves that $(f_n)_{n\in\mathbb{N}}$ converges to f in $L^1(J;V)$.

Exercise 64.4 ($L^q((0,1);L^p(0,1))$). (i) g is almost separably valued since $V := L^p(D)$ is separable for all $p\in[1,\infty)$ (see e.g., Brezis [6, Thm. 4.13].
(ii) Identifying $(L^p(D))'$ with $L^{p'}(D)$ (with the convention that $p'=\infty$ for $p=1$; see, e.g., [6, Thm. 4.11]), we have

$$\langle w, g(t)\rangle_{(L^p)',L^p} = \int_D w(x)g(t)(x)\,\mathrm{d}x = \int_0^t w(x)\,\mathrm{d}x,$$

for all $w\in(L^p(D))' = L^{p'}(D)$. The function $J\ni t\mapsto \int_0^t w(x)\,\mathrm{d}x$ is measurable since it is continuous. Hence, g is weakly measurable.

(iii) We infer from Pettis measurability theorem (Theorem 64.4) that g is strongly measurable. To conclude, we have to prove that $\int_0^1 \|g(t)\|_V^q \mathrm{d}t < \infty$ if $q \in [1,\infty)$ and $\operatorname{ess\,sup}_{t\in(0,1)} \|g(t)\|_V < \infty$ if $q = \infty$. We have $\|g(t)\|_V = \|g(t)\|_{L^p(0,1)} = t^{\frac{1}{p}}$, and for all $q \in [1,\infty)$, we have

$$\int_J \|g(t)\|_V^q \mathrm{d}t = \int_0^1 t^{\frac{q}{p}} \mathrm{d}t = \frac{p}{q+p}.$$

Hence, $\|g\|_{L^q(J;V)} = (\frac{p}{q+p})^{\frac{1}{q}} < \infty$. We also have $\|g\|_{L^\infty(J;V)} = 1 < \infty$.

Exercise 64.5 (Constants). We follow the proof of Lemma 2.11. Let $\rho \in C_0^\infty(J;\mathbb{R})$ be s.t. $\int_J \rho \,\mathrm{d}x = 1$, and set $c_\rho := \int_J f(\xi)\rho(\xi)\,\mathrm{d}\xi \in V$. Let φ be an arbitrary function in $C_0^\infty(J;\mathbb{R})$ and set $c_\varphi := \int_J \varphi(\xi)\,\mathrm{d}\xi$. The function $\psi(t) := \int_0^t (\varphi(\xi) - c_\varphi\rho(\xi))\,\mathrm{d}\xi$ is in $C_0^\infty(J;\mathbb{R})$ by construction, and we have $\partial_t\psi(t) = \varphi(t) - c_\varphi\rho(t)$. Since $\int_J f(t)\partial_t\psi(t)\mathrm{d}t = -\int_J (\partial_t f(t))\psi(t)\mathrm{d}t = 0$ by assumption, we infer that

$$\int_J f(t)\varphi(t)\mathrm{d}t = \int_J f(t)(\partial_t\psi(t) + c_\varphi\rho(t))\mathrm{d}t = c_\varphi \int_J f(t)\rho(t)\mathrm{d}t = c_\rho \int_J \varphi(t)\,\mathrm{d}x.$$

Hence, $\int_J (f(t) - c_\rho)\varphi(t)\mathrm{d}t = 0$ for all $\varphi \in C_0^\infty(J;\mathbb{R})$. Corollary 64.28 shows that $f = c_\rho$.

Exercise 64.6 (Linear map). Let $v \in L^1_{\mathrm{loc}}(J;V)$ and assume that v is weakly differentiable in $L^1_{\mathrm{loc}}(J;V)$.

Let us first verify that $K(v) \in L^1_{\mathrm{loc}}(J;W)$. Let J_0 be a compact subset of J. Then $\mathbb{1}_{J_0} v \in L^1(J;V)$, and Corollary 64.14 implies that $K(\mathbb{1}_{J_0} v) \in L^1(J;W)$ which proves that $K(v)$ is integrable on J_0 since $K(v)(t) = K(v(t)) = K(\mathbb{1}_{J_0}(t)v(t)) = K(\mathbb{1}_{J_0}v)(t)$ for a.e. $t \in J_0$.

Let us now prove that $K(v)$ is weakly differentiable with $\partial_t(K(v)) = K(\partial_t v)$. Let $\phi \in C_0^\infty(J)$. We have $\phi\partial_t v \in L^1(J;V)$ because $\partial_t v \in L^1_{\mathrm{loc}}(J;V)$ and ϕ is compactly supported in J. Owing to Corollary 64.14, we infer that $K(\phi\partial_t v) \in L^1(J;W)$. But the linearity of K implies that $\phi(t)K(\partial_t v(t)) = K(\phi(t)\partial_t v(t)) = K(\phi\partial_t v)(t)$ for a.e. $t \in J$. Hence, we have

$$\begin{aligned}\int_J \phi(t)K(\partial_t v(t))\mathrm{d}t &= \int_J K(\phi(t)\partial_t v)(t)\mathrm{d}t = K\Big(\int_J \phi(t)\partial_t v(t)\mathrm{d}t\Big)\\ &= K\Big(-\int_J v(t)\partial_t\phi(t)\mathrm{d}t\Big) = -\int_J K(v(t)\partial_t\phi(t))\mathrm{d}t\\ &= -\int_J K(v(t))\partial_t\phi(t)\mathrm{d}t.\end{aligned}$$

This proves that $K(v)$ is weakly differentiable with $\partial_t(K(v)) = K(\partial_t v)$.

Exercise 64.7 ($X^{p,q}(J;V,W)$). Let us consider a Cauchy sequence $(v_n)_{n\in\mathbb{N}}$ in $X^{p,q}(J;V,W)$. Then $v_n \to v$ in $L^p(J;V)$ and $\partial_t v_n \to w \in L^q(J;W)$. Let $\phi \in C_0^\infty(J;\mathbb{R})$. We have $\int_J \phi(t)v_n(t)\mathrm{d}t \to \int_J \phi(t)v(t)\mathrm{d}t$ in V since

$$\begin{aligned}\Big\|\int_J \phi(t)(v_n(t) - v(t))\mathrm{d}t\Big\|_V &\le \|\phi\|_{C^0(\overline{J};\mathbb{R})}\|v_n - v\|_{L^1(J;V)}\\ &\le \|\phi\|_{C^0(\overline{J};\mathbb{R})}T^{1-\frac{1}{p}}\|v_n - v\|_{L^p(J;V)}.\end{aligned}$$

As a result, we have

$$-\int_0^T \phi'(t)v(t)\mathrm{d}t \leftarrow -\int_0^T \phi'(t)v_n(t)\mathrm{d}t = \int_0^T \phi(t)\partial_t v_n(t)\mathrm{d}t \to \int_0^T \phi(t)w(t)\mathrm{d}t.$$

We conclude that $\partial_t v = w$, so that $v \in X^{p,q}(J;V,W)$.

Exercise 64.8 (Continuous embedding). (i) Observe first that the definitions of v and w imply that $v \in L^q(J;W)$ and $w(t) \in W$ for all $t \in \overline{J}$ since $\|w(t)\|_W \le \|v\|_{L^1(J;W)} \le T^{\frac{q-1}{q}}\|v\|_{L^q(J;W)}$. Let $t \in \overline{J}$ and $t_n \in \overline{J}$ be such that $t_n \to t$ as $n \to \infty$. Let J_n be the interval (t, t_n) or (t_n, t), and let $\mathbb{1}_{J_n}$ be the indicator function of J_n. We have

$$\|w(t_n) - w(t)\|_W = \left\| \int_t^{t_n} v(\tau)\,\mathrm{d}\tau \right\|_W = \int_J \mathbb{1}_{J_n}\|v(\tau)\|_W\,\mathrm{d}\tau.$$

The sequence $\mathbb{1}_{J_n}(\tau)\|v(\tau)\|_W$ converges a.e. to 0 and $\|\mathbb{1}_{J_n}v\|_W \le \|v\|_W$. Hence, Lebesgue's dominated convergence theorem implies that $\|w(t_n) - w(t)\|_W \to 0$ as $t_n \to t$, thereby proving that $w \in C^0(\overline{J};W)$. Notice in passing that if $q > 1$, we also have

$$\begin{aligned}\|w(s) - w(t)\|_W = \left\| \int_t^s v(\tau)\,\mathrm{d}\tau \right\|_W &\le \left(\int_t^s \mathrm{d}\tau \right)^{\frac{q-1}{q}} \|v\|_{L^q(J;W)} \\ &\le |s-t|^{1-\frac{1}{q}}\|v\|_{L^q(J;W)},\end{aligned}$$

which shows that w in $C^{0,1-\frac{1}{q}}(\overline{J};W)$.
(ii) Let us evaluate $\lim_{n\to\infty}\phi_n(\tau)$. Let $\mathbb{1}_{(0,\tau)}$ be the indicator function of $(0,\tau)$. We have

$$\phi_n(\tau) = \int_{-\infty}^{+\infty} (\rho_n(s-\xi) - \rho_n(t-\xi))\mathbb{1}_{(0,\tau)}(\xi)\,\mathrm{d}\xi.$$

It is a standard result about mollifiers that $\int_{-\infty}^{+\infty}\rho_n(s-\xi)f(\xi)\,\mathrm{d}\xi$ converges to $f(s)$ for a.e. s and all $f \in L^1(\mathbb{R})$. Hence, we have

$$\lim_{n\to\infty}\phi_n(\tau) = \mathbb{1}_{(0,\tau)}(s) - \mathbb{1}_{(0,\tau)}(t) = \mathbb{1}_{(s,t)}(\tau), \quad \forall \tau \notin \{s,t\}.$$

(iii) Observe first that the definition of $\delta_n(s,t)$ makes sense since $0 \le s - \frac{y}{n} \le t - \frac{y}{n} \le T$ for all $y \in [-1,1]$, because we have assumed that $n \ge N \ge \max(\frac{1}{s}, \frac{1}{T-t})$ and $0 < s < t < T$. Up to two changes of variable, we have

$$\begin{aligned}\delta_n(s,t) &= \int_{-1}^1 \rho(y)(u(s-\frac{y}{n}) - u(t-\frac{y}{n}))\,\mathrm{d}y \\ &= \int_J (\rho_n(s-z) - \rho_n(t-z))u(z)\,\mathrm{d}z.\end{aligned}$$

Since $\phi_n'(\tau) = \rho_n(s-\tau) - \rho_n(t-\tau)$, we infer that

$$\delta_n(s,t) = \int_J \phi_n'(\tau)u(\tau)\,\mathrm{d}\tau.$$

Notice that ϕ_n is in $C_0^\infty(\mathbb{R};\mathbb{R})$. By definition of $v(\tau) := \partial_t u(\tau)$ (see Definition 64.29), we have

$$\delta_n(s,t) = \int_J \phi_n'(\tau)u(\tau)\,\mathrm{d}\tau := -\int_J \phi_n(\tau)v(\tau)\,\mathrm{d}\tau.$$

(iv) We observe that

$$\left\| \delta_n(s,t) + \int_s^t v(\tau)\,\mathrm{d}\tau \right\|_W \le \int_J |\phi_n(\tau) - \mathbb{1}_{(s,t)}(\tau)|\|v(\tau)\|_W\,\mathrm{d}\tau.$$

We now apply Lebesgue's dominated convergence theorem and conclude that

$$\lim_{n\to\infty} \delta_n(s,t) = -\int_s^t v(\tau)\,\mathrm{d}\tau = w(s) - w(t).$$

Notice that we also have

$$\lim_{n\to\infty} \delta_n(s,t) = u(s) - u(t),$$

since $\int_{-1}^1 \rho(y)u(s-\frac{y}{n})\,\mathrm{d}y \to u(s)$ as $n\to\infty$ (recall that $u \in L^p(J;V) \hookrightarrow L^1(J;V)$).
(v) The above argument shows that w and u differ by a constant, i.e., there is $a \in W$ such that $u = w + a$ a.e. on J. This proves that $u \in C^0(\overline{J};W)$ since we have already established that $w \in C^0(\overline{J};W)$. Actually, we have also established that $u \in C^{0,\frac{q-1}{q}}(\overline{J};W)$ if $q > 1$.

Exercise 64.9 (Time derivative of product). Let $\phi \in C_0^\infty(J;\mathbb{R})$. Observing that $\alpha u \in L^1_{\mathrm{loc}}(J;V)$, we have

$$\int_J u(t)\alpha(t)\partial_t\phi(t)\mathrm{d}t = \int_J u(t)\partial_t(\alpha(t)\phi(t))\mathrm{d}t - \int_J u(t)\phi(t)\partial_t\alpha(t)\mathrm{d}t.$$

Since $C_0^\infty(J)$ is dense in $C_0^1(J)$, we can apply (64.3) for every test function in $C_0^1(J)$ (just apply Lebesgue's dominated convergence theorem). By abusing the notation and identifying u with its image by the canonical injection mapping from V to W, we have

$$\begin{aligned}\int_J u(t)\alpha(t)\partial_t\phi(t)\mathrm{d}t &= -\int_J (\partial_t u(t))\alpha(t)\phi(t)\mathrm{d}t - \int_J u(t)\phi(t)\partial_t\alpha(t)\mathrm{d}t\\ &= -\int_J (\alpha(t)\partial_t u(t) + u(t)\partial_t\alpha(t))\phi(t)\mathrm{d}t,\end{aligned}$$

which proves that $\partial_t(\alpha u) = u\partial_t\alpha + \alpha\partial_t u$.

Chapter 65

Weak formulation and well-posedness

Exercises

Exercise 65.1 (L^p-integrability of $A(u)$). Let $u \in L^p(J;V)$ and let $A(u)$ be defined in (65.6). Prove that $A(u) \in L^p(J;V')$ with $\|A(u)\|_{L^p(J;V')} \le M\|u\|_{L^p(J;V)}$. (*Hint*: use Theorem 64.12.)

Exercise 65.2 (Ultraweak formulation). Write the ultraweak formulation for the heat equation.

Exercise 65.3 (Gronwall's lemma). Let $J := (0,T)$, $T > 0$. Let $\alpha, \beta, u \in L^1(J;\mathbb{R})$ be s.t. $\alpha\beta, \beta u \in L^1(J;\mathbb{R})$, $\beta(t) \ge 0$, and $u(t) \le \alpha(t) + \int_0^t \beta(r)u(r)\,\mathrm{d}r$ for a.e. $t \in J$. (i) Prove that $v(t) := e^{-\int_0^t \beta(r)\,\mathrm{d}r}\int_0^t \beta(r)u(r)\,\mathrm{d}r$ is in $W^{1,1}(J;\mathbb{R})$. (ii) Prove that $v(t) \le \int_0^t \alpha(r)\beta(r)e^{-\int_0^r \beta(s)\,\mathrm{d}s}\,\mathrm{d}r$. (iii) Prove that

$$u(t) \le \alpha(t) + \int_0^t \alpha(s)\beta(s)e^{\int_s^t \beta(r)\,\mathrm{d}r}\,\mathrm{d}s. \tag{65.1}$$

(*Hint*: use Step (ii) and $\int_0^t \beta(r)u(r)\,\mathrm{d}r = v(t)e^{\int_0^t \beta(r)\,\mathrm{d}r}$.) (iv) Assume now that α is nondecreasing, i.e., $\alpha(r) \le \alpha(t)$ for a.e. $r,t \in J$ s.t. $r \le t$. Prove that for a.e. $t \in J$,

$$u(t) \le \alpha(t)e^{\int_0^t \beta(r)\,\mathrm{d}r}. \tag{65.2}$$

(v) Assume that β is constant and $\alpha \in W^{1,1}(J)$. Prove that for a.e. $t \in J$, $u(t) \le \alpha(0)e^{\beta t} + \int_0^t \alpha'(r)e^{\beta(t-r)}\,\mathrm{d}r$. *Note*: owing to the assumption $\beta(t) \ge 0$, Gronwall's lemma can be used to show that the function u has at most exponential growth in time, but it cannot be used to show that u has exponential decay. However, if the assumption $u(t) \le \alpha(t) + \int_0^t \beta(r)u(r)\,\mathrm{d}r$ is replaced by the stronger assumption $u'(t) \le \alpha'(t) + \beta(t)u(t)$, then $u(t) \le e^{\int_0^t \beta(r)\,\mathrm{d}r}u(0) + \int_0^t \alpha'(r)e^{\int_r^t \beta(s)\,\mathrm{d}s}\,\mathrm{d}r$ regardless of the sign of β.

Exercise 65.4 (Exponentially decaying estimate). (i) Prove the a priori estimate (65.17). (*Hint*: adapt the proof of Lemma 65.10 by considering the test function $(0,w) \in Y$ with $w(t) := e^{2\frac{t}{\rho}}u(t)$ and the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$.) (ii) Assuming that $f \in L^\infty((0,\infty);V')$, prove that $\limsup_{t\to\infty}\|u(t)\|_L \le \frac{\iota_{L,V}}{\alpha}\|f\|_{L^\infty((0,\infty);V')}$. (*Hint*: use (65.17).)

Solution to exercises

Exercise 65.1 (L^p-integrability of $A(u)$). Assume first that $p \in [1,\infty)$. Owing to (65.5b), we infer the bound

$$\left(\int_J \|A(t)(u(t))\|_{V'}^p \mathrm{d}t\right)^{\frac{1}{p}} \le M\left(\int_J (u(t))\|_V^p \mathrm{d}t\right)^{\frac{1}{p}} = M\|u\|_{L^p(J;V)}.$$

Bochner's theorem (Theorem 64.12) implies that $A(u) \in L^p(J;V')$ since $A(u)$ is strongly measurable. A similar argument applies if $p=\infty$.

Exercise 65.2 (Ultraweak formulation). The ultraweak formulation for the heat equation leads to the trial space

$$X_{\mathrm{uw}} := L^2(J;H_0^1(D))$$

and to the test space

$$Y_{\mathrm{uw}} := \{w \in L^2(J;H_0^1(D)) \mid \partial_t w \in L^2(J;H^{-1}(D)),\ w(T)=0\}.$$

The corresponding forms are

$$b_{\mathrm{uw}}(v,w) := \int_J \Big(\langle v(t), -\partial_t w(t)\rangle + (\kappa(t)\nabla v(t), \nabla w(t))_{\boldsymbol{L}^2(D)}\Big)\mathrm{d}t,$$
$$\ell_{\mathrm{uw}}(w) := (u_0, w(0)_{L^2(D)} + \int_J \langle f(t), w(t)\rangle \mathrm{d}t.$$

Exercise 65.3 (Gronwall's lemma). (i) Let us consider the function

$$v(t) := e^{-\int_0^t \beta(r)\,\mathrm{d}r} \int_0^t \beta(r)u(r)\,\mathrm{d}r.$$

The assumptions imply that v is continuous on $\overline{J}$. The weak derivative of v is

$$v'(t) = \beta(t)\left(u(t) - \int_0^t \beta(r)u(r)\,\mathrm{d}r\right) e^{-\int_0^t \beta(r)\,\mathrm{d}r}.$$

The assumptions imply that $v' \in L^1(J;\mathbb{R})$ (notice that $\beta u \in L^1(J;\mathbb{R})$), thereby showing that $v \in W^{1,1}(J;\mathbb{R})$.
(ii) Observing that $v(0)=0$, the above computation shows that

$$v(t) = \int_0^t \beta(r)\left(u(r) - \int_0^r \beta(s)u(s)\,\mathrm{d}s\right) e^{-\int_0^r \beta(s)\,\mathrm{d}s}\,\mathrm{d}r,$$

which, in turn, implies that

$$v(t) \le \int_0^t \alpha(r)\beta(r) e^{-\int_0^r \beta(s)\,\mathrm{d}s}\,\mathrm{d}r,$$

since $\beta(t) \ge 0$ for a.e. $t \in J$.
(iii) We follow the hint. The definition of v implies that

$$\int_0^t \beta(r)u(r)\,\mathrm{d}r = v(t)e^{\int_0^t \beta(r)\,\mathrm{d}r},$$

and Step (ii) implies that

$$\begin{aligned}\int_0^t \beta(r)u(r)\,\mathrm{d}r &\le e^{\int_0^t \beta(r)\,\mathrm{d}r}\int_0^t \alpha(r)\beta(r)e^{-\int_0^r \beta(s)\,\mathrm{d}s}\,\mathrm{d}r\\ &= \int_0^t \alpha(r)\beta(r)e^{-\int_t^r \beta(s)\,\mathrm{d}s}\,\mathrm{d}r,\end{aligned}$$

observing that $\alpha\beta \in L^1(J;\mathbb{R})$ by assumption. This, in turn, implies the expected inequality: For a.e. $t \in J$,

$$u(t) \le \alpha(t) + \int_0^t \alpha(r)\beta(r)e^{\int_r^t \beta(s)\,\mathrm{d}s}\,\mathrm{d}r.$$

(iv) Assume now that α is nondecreasing, i.e., $\alpha(r) \le \alpha(t)$ for a.e. $r, t \in J$ s.t. $r \le t$. The above inequality implies that

$$\begin{aligned}u(t) &\le \alpha(t)\left(1 + \int_0^t \beta(r)e^{\int_r^t \beta(s)\,\mathrm{d}s}\,\mathrm{d}r\right)\\ &= \alpha(t)\left(1 - \int_0^t \frac{\mathrm{d}}{\mathrm{d}r}\Big(\int_r^t \beta(s)\,\mathrm{d}s\Big)e^{\int_r^t \beta(s)\,\mathrm{d}s}\,\mathrm{d}r\right)\\ &= \alpha(t)\left(1 - \int_0^t \frac{\mathrm{d}}{\mathrm{d}r}\Big(e^{\int_r^t \beta(s)\,\mathrm{d}s}\Big)\,\mathrm{d}r\right).\end{aligned}$$

We can now conclude that for a.e. $t \in J$, we have

$$u(t) \le \alpha(t)e^{\int_0^t \beta(r)\,\mathrm{d}r}.$$

(v) Let us apply Step (iii) assuming that β is constant and $\alpha \in W^{1,1}(J)$. We obtain for a.e. $t \in J$,

$$\begin{aligned}u(t) &\le \alpha(t) + \int_0^t \alpha(r)\beta e^{\beta(t-r)}\,\mathrm{d}r\\ &= \alpha(t) - \int_0^t \alpha(r)\frac{\mathrm{d}}{\mathrm{d}r}e^{\beta(t-r)}\,\mathrm{d}r\\ &= \alpha(t) - \alpha(t) + \alpha(0)e^{\beta t} + \int_0^t \alpha'(r)e^{\beta(t-r)}\,\mathrm{d}r\\ &= \alpha(0)e^{\beta t} + \int_0^t \alpha'(r)e^{\beta(t-r)}\,\mathrm{d}r.\end{aligned}$$

Exercise 65.4 (Exponentially decaying estimate). (i) Let $t \in (0,T]$. Following the hint, let us consider the function $w \in L^2(J;V)$ s.t. $w(t) := e^{2\frac{t}{\rho}}u(t)$. We first observe that $\partial_t w(\tau) := \frac{2}{\rho}e^{2\frac{\tau}{\rho}}u(\tau) + e^{2\frac{\tau}{\rho}}\partial_t u(\tau)$ for all $\tau \in (0,t)$. Invoking Lemma 64.40 (integration by parts in time over the interval $(0,t)$), we infer that

$$\int_0^t \langle \partial_t u(\tau), w(\tau)\rangle_{V',V}\,\mathrm{d}\tau = -\frac{1}{\rho}\int_0^t e^{2\frac{\tau}{\rho}}\|u(\tau)\|_L^2\,\mathrm{d}\tau + \frac{1}{2}e^{2\frac{t}{\rho}}\|u(t)\|_L^2 - \frac{1}{2}\|u_0\|_L^2.$$

Using the boundedness of the embedding $V \hookrightarrow L$ and the definition of ρ gives

$$\frac{1}{\rho}\|u(\tau)\|_L^2 = \frac{\alpha}{2\iota_{L,V}^2}\|u(\tau)\|_L^2 \le \frac{\alpha}{2}\|u(\tau)\|_V^2.$$

Hence, we have

$$\int_0^t \langle \partial_t u(\tau), w(\tau)\rangle_{V',V}\,\mathrm{d}\tau \geq -\frac{\alpha}{2}\int_0^t e^{2\frac{\tau}{\rho}}\|u(\tau)\|_V^2\,\mathrm{d}\tau + \frac{1}{2}e^{2\frac{t}{\rho}}\|u(t)\|_L^2 - \frac{1}{2}\|u_0\|_L^2.$$

The coercivity property (65.5c) implies that

$$\begin{aligned}
&\frac{1}{2}e^{2\frac{t}{\rho}}\|u(t)\|_L^2 - \frac{1}{2}\|u_0\|_L^2 + \frac{1}{2}\alpha\int_0^t e^{2\frac{\tau}{\rho}}\|u(\tau)\|_V^2\,\mathrm{d}\tau\\
&\leq \int_0^t \langle \partial_t u(\tau), w(\tau)\rangle_{V',V}\,\mathrm{d}\tau + \alpha\int_0^t e^{2\frac{\tau}{\rho}}\|u(\tau)\|_V^2\,\mathrm{d}\tau\\
&\leq \int_0^t \langle \partial_t u(\tau), w(\tau)\rangle_{V',V}\,\mathrm{d}\tau + \int_0^t e^{2\frac{\tau}{\rho}}\langle A(\tau)(u(\tau)), u(\tau)\rangle_{V',V}\,\mathrm{d}\tau\\
&= \int_0^t \langle \partial_t u(\tau), w(\tau)\rangle_{V',V}\,\mathrm{d}\tau + \int_0^t \langle A(\tau)(u(\tau)), w(\tau)\rangle_{V',V}\,\mathrm{d}\tau.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\int_0^t \langle \partial_t u(\tau) + A(\tau)(u(\tau)), w(\tau)\rangle_{V',V}\,\mathrm{d}\tau &= b(u,(0,w)) = \ell((0,w))\\
&= \int_0^t \langle f(\tau), w(\tau)\rangle_{V',V}\,\mathrm{d}\tau\\
&\leq \int_0^t e^{2\frac{\tau}{\rho}}\|f(\tau)\|_{V'}\|u(\tau)\|_V\,\mathrm{d}\tau\\
&\leq \int_0^t e^{2\frac{\tau}{\rho}}\big(\frac{1}{2}\alpha\|u(\tau)\|_V^2 + \frac{1}{2\alpha}\|f(\tau)\|_{V'}^2\big)\,\mathrm{d}\tau\\
&= \frac{1}{2}\alpha\int_0^t e^{2\frac{\tau}{\rho}}\|u(\tau)\|_V^2\,\mathrm{d}\tau + \frac{1}{2\alpha}\int_0^t \|e^{\frac{\tau}{\rho}}f(\tau)\|_{V'}^2\,\mathrm{d}\tau,
\end{aligned}$$

where we used Young's inequality in the last bound. Putting everything together, we conclude that

$$\frac{1}{2}e^{2\frac{t}{\rho}}\|u(t)\|_L^2 - \frac{1}{2}\|u_0\|_L^2 \leq \frac{1}{2\alpha}\int_0^t \|e^{\frac{\tau}{\rho}}f(\tau)\|_{V'}^2\,\mathrm{d}\tau,$$

and rearranging the terms leads to the a priori estimate (65.17).
(ii) Since $f \in L^\infty((0,\infty);V')$, taking square roots in the a priori estimate (65.17) and recalling that $J_t := (0,t)$, we infer that

$$\begin{aligned}
\|u(t)\|_L &\leq e^{-\frac{t}{\rho}}\|u_0\|_L + \alpha^{-\frac{1}{2}}\|e^{-\frac{t-\cdot}{\rho}}f\|_{L^2(J_t;V')}\\
&\leq e^{-\frac{t}{\rho}}\|u_0\|_L + \alpha^{-\frac{1}{2}}\|e^{-2\frac{t-\cdot}{\rho}}\|_{L^1(J_t)}^{\frac{1}{2}}\|f\|_{L^\infty(J_t;V')}\\
&\leq e^{-\frac{t}{\rho}}\|u_0\|_L + \frac{\iota_{L,V}}{\alpha}\|f\|_{L^\infty(J_t;V')},
\end{aligned}$$

since $\|e^{-2\frac{t-\cdot}{\rho}}\|_{L^1(J_t)} = \int_0^t e^{-2\frac{t-s}{\rho}}\,\mathrm{d}s = \frac{\rho}{2}(1-e^{-2\frac{t}{\rho}}) \leq \frac{\rho}{2} = \frac{\iota_{L,V}^2}{\alpha}$. The conclusion is straightforward since $\lim_{t\to\infty} e^{-\frac{t}{\rho}} = 0$.

Chapter 66

Semi-discretization in space

Exercises

Exercise 66.1 ($L^2(J;V)$-estimate using elliptic projection). Use the notation from §66.3.1. Assume that the elliptic projection is time-independent and set $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(u(t))$ for all $t \in J$. Prove that

$$\|u - u_h\|_{L^2(J;V)} \le \|\eta\|_{L^2(J;V)} + \frac{1}{\alpha}\|\partial_t \eta\|_{L^2(J;V')} + \frac{2}{\sqrt{\alpha}}\|\eta(0)\|_L.$$

(*Hint*: use the error equation (66.12).)

Exercise 66.2 (Naive $C^0(\overline{J};L)$-estimate). Use the proof of Theorem 66.7 to derive an upper bound on $\|u - u_h\|_{C^0(\overline{J};L)}$. (*Hint*: integrate (66.10) in time over the interval $J_s := (0,s)$ for all $s \in (0,T]$.) Assuming smoothness, is the convergence rate of this error estimate optimal for the heat equation? What is the term that limits the convergence rate?

Exercise 66.3 (Theorem 66.9). Prove the error estimate (66.15). (*Hint*: see Exercise 65.4.)

Exercise 66.4 (Lemma 66.17). Let $\Pi_h^{\mathrm{E}}(t) \in \mathcal{L}(H_0^1(D);V_h)$ be defined in (66.11) for the time-dependent heat equation. Let $u \in H^1(J;H_0^1(D))$ and set $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(t;u(t))$ for a.e. $t \in J$. (i) Prove that

$$|\partial_t \eta(t)|_{H^1(D)} \le |\partial_t u(t) - \Pi_h^{\mathrm{E}}(t;\partial_t u(t))|_{H^1(D)} + \rho^{-1}\frac{M'}{\alpha}|\eta(t)|_{H^1(D)}.$$

(ii) Prove (66.21). (*Hint*: use the adjoint problem $a(t;v,\xi(t)) = (\delta_h(t),v)_{L^2(D)}$ for all $v \in H_0^1(D)$, with $\delta_h(t) := \partial_t(\Pi_h^{\mathrm{E}}(t;u(t))) - \Pi_h^{\mathrm{E}}(t;\partial_t u(t))$ for a.e. $t \in J$, and show that

$$\|\delta_h(t)\|_{L^2(D)}^2 = a(t;\delta_h(t),\xi(t) - w_h) + \dot{a}(t;\eta(t),w_h - \xi(t)) + \dot{a}(t;\eta(t),\xi(t)),$$

for all $w_h \in V_h$.) (iii) Show that $\|\Pi_h^{\mathrm{E}}(t;u(t))\|_{H^1(J;V_h)} \le c(\alpha,M,\frac{M'}{\rho})\|u\|_{H^1(J;V)}$ for all $u \in C^\infty(\overline{J};V)$ and all $h \in \mathcal{H}$.

Solution to exercises

Exercise 66.1 ($L^2(J;V)$-estimate using elliptic projection). Using the test function $w_h := e_h(t)$ for all $t \in J$ in the error equation (66.12), integrating over time, and invoking Young's inequality gives

$$\alpha\|e_h\|^2_{L^2(J;V)} \le \frac{1}{\alpha}\|\partial_t\eta\|^2_{L^2(J;V')} + \|e_h(0)\|^2_L,$$

where we dropped $\|e_h(T)\|^2_L$ on the left-hand side. Dividing by α, taking the square root, using that $\|e_h(0)\|_L \le 2\|\eta(0)\|_L$, and invoking the triangle inequality yields the assertion.

Exercise 66.2 (Naive $C^0(\overline{J};L)$-estimate). Integrating (66.10) in time over the interval $J_s := (0,s)$ for all $s \in (0,T)$, we infer that

$$\|e_h(s)\|^2_L \le \frac{1}{\alpha}\|\partial_t\eta + A(t)\eta\|^2_{L^2(J_s;V')} + \|e_h(0)\|^2_L,$$

where we dropped the term $\alpha\|e_h\|^2_{L^2(J_s;V)}$ on the left-hand side. We now bound the right-hand side by replacing J_s by the full time interval J, then we exploit that s is arbitrary in $(0,T]$ on the left hand-side (and the bound for $s=0$ is obvious). We infer that

$$\|e_h\|^2_{C^0(\overline{J};L)} \le \frac{1}{\alpha}\|\partial_t\eta + A(t)\eta\|^2_{L^2(J;V')} + \|e_h(0)\|^2_L.$$

Taking the square root, observing that $\|\partial_t\eta + A(t)\eta\|_{L^2(J;V')} \le \|\partial_t\eta\|_{L^2(J;V')} + M\|\eta\|_{L^2(J;V)}$ and since $\|e_h(0)\|_L \le \|\eta(0)\|_L$, we infer that

$$\frac{1}{\sqrt{\alpha}}\|e_h\|_{C^0(\overline{J};L)} \le \frac{1}{\alpha}\|\partial_t\eta\|_{L^2(J;V')} + \frac{M}{\alpha}\|\eta\|_{L^2(J;V)} + \frac{1}{\sqrt{\alpha}}\|\eta(0)\|_L.$$

Invoking the triangle inequality for $u - u_h = \eta - e_h$ shows that

$$\frac{1}{\sqrt{\alpha}}\|u-u_h\|_{C^0(\overline{J};L)} \le \frac{1}{\alpha}\|\partial_t\eta\|_{L^2(J;V')} + \frac{M}{\alpha}\|\eta\|_{L^2(J;V)} + \frac{2}{\sqrt{\alpha}}\|\eta\|_{C^0(\overline{J};L)}.$$

The second term on the right-hand side is the one that gives a convergence rate that is not optimal for smooth solutions of the heat equation. Indeed, this term typically decays as $\mathcal{O}(h^r)$, whereas the other terms on the right-hand side decay as $\mathcal{O}(h^{r+1})$.

Exercise 66.3 (Theorem 66.9). We take $v_h(t) := \Pi^{\mathrm{E}}_h(u(t))$ for all $t \in J$, i.e., we work with the error equation (66.12). Let us consider the test function $w_h(t) := e^{2\frac{t}{\rho}}e_h(t)$ for all $t \in J$. Then we can proceed as in Exercise 65.4 and invoke exactly the same arguments, leading to the bound

$$\|e_h(t)\|_L \le \frac{1}{\sqrt{\alpha}}\|e^{-\frac{t-\cdot}{\rho}}\partial_t\eta\|_{L^2(J_t;V')} + e^{-\frac{t}{\rho}}\|e_h(0)\|_L.$$

We conclude by invoking the bound $\|e_h(0)\|_L \le \|\eta(0)\|_L$ and the triangle inequality on $u - u_h = \eta - e_h$.

Exercise 66.4 (Lemma 66.17). (i) Recalling (66.19), and using the coercivity of a and the boundedness of $\dot{a}$, we infer that

$$\begin{aligned}\alpha\,|\partial_t(\Pi^{\mathrm{E}}_h(t;u(t))) - \Pi^{\mathrm{E}}_h(t;\partial_t u(t))|_{H^1(D)} &\le \sup_{w_h\in V_h}\frac{|a(t;\partial_t(\Pi^{\mathrm{E}}_h(t;u(t))) - \Pi^{\mathrm{E}}_h(t;\partial_t u(t)),w_h)|}{|w_h|_{H^1(D)}}\\ &= \sup_{w_h\in V_h}\frac{|\dot{a}(t;\eta(t),w_h)|}{|w_h|_{H^1(D)}} \le \rho^{-1}M'|\eta(t)|_{H^1(D)}.\end{aligned}$$

Dividing by α and invoking the triangle inequality for

$$\partial_t\eta(t) = (\partial_t u(t) - \Pi_h^{\mathrm{E}}(t;\partial_t u(t))) - (\partial_t(\Pi_h^{\mathrm{E}}(t;u(t))) - \Pi_h^{\mathrm{E}}(t;\partial_t u(t)))$$

proves the claim.
(ii) Let us set $\delta_h(t) := \partial_t(\Pi_h^{\mathrm{E}}(t;u(t))) - \Pi_h^{\mathrm{E}}(t;\partial_t u(t))$. Considering the dual problem suggested in the hint, we infer that

$$\begin{aligned}\|\delta_h(t)\|_{L^2(D)}^2 &= a(t;\delta_h(t),\xi(t))\\ &= a(t;\delta_h(t),\xi(t)-w_h) + a(t;\delta_h(t),w_h)\\ &= a(t;\delta_h(t),\xi(t)-w_h) + \dot a(t;\eta(t),w_h)\\ &= a(t;\delta_h(t),\xi(t)-w_h) + \dot a(t;\eta(t),w_h-\xi(t)) + \dot a(t;\eta(t),\xi(t)),\end{aligned}$$

for all $w_h \in V_h$, where we used (66.19) in the third line. Invoking the boundedness of a for the first term, that of $\dot a$ for the second term, and (66.20) for the third term, we infer that

$$\begin{aligned}\|\delta_h(t)\|_{L^2(D)}^2 \le \big(M|\delta_h(t)|_{H^1(D)} + \rho^{-1}M'|\eta(t)|_{H^1(D)}\big)|\xi(t)-w_h|_{H^1(D)}\\ + \rho^{-1}M''|\eta(t)|_{H^{1-s}(D)}|\xi(t)|_{H^{1+s}(D)}.\end{aligned}$$

Taking the infimum over $w_h \in V_h$ and using the approximation properties of finite elements, we obtain

$$\begin{aligned}\|\delta_h(t)\|_{L^2(D)}^2 \le \Big(c\,h^s\big(M|\delta_h(t)|_{H^1(D)} + \rho^{-1}M'|\eta(t)|_{H^1(D)}\big)\\ + \rho^{-1}M''|\eta(t)|_{H^{1-s}(D)}\Big)|\xi(t)|_{H^{1+s}(D)}.\end{aligned}$$

The elliptic regularity property $|\xi(t)|_{H^{1+s}(D)} \le c_{\mathrm{smo}}\alpha^{-1}\ell_D^{1-s}\|\delta_h(t)\|_{L^2(D)}$ leads to

$$\begin{aligned}\|\delta_h(t)\|_{L^2(D)} \le c\,\alpha^{-1}\ell_D^{1-s}\Big(h^s\big(M|\delta_h(t)|_{H^1(D)} + \rho^{-1}M'|\eta(t)|_{H^1(D)}\big)\\ + \rho^{-1}M''|\eta(t)|_{H^{1-s}(D)}\Big).\end{aligned}$$

Since $|\delta_h(t)|_{H^1(D)} \le \rho^{-1}\frac{M'}{\alpha}|\eta(t)|_{H^1(D)}$ as established in Step (i), we infer that

$$\|\delta_h(t)\|_{L^2(D)} \le c\,\rho^{-1}\alpha^{-1}\ell_D^{1-s}\left(h^s\Big(1+\frac{M}{\alpha}\Big)M'|\eta(t)|_{H^1(D)}\right) + M''|\eta(t)|_{H^{1-s}(D)}\Big).$$

Since $\|\eta(t)\|_{L^2(D)} \le c\frac{M}{\alpha}h^s\ell_D^{1-s}|\eta(t)|_{H^1(D)}$, the Riesz–Thorin theorem (Theorem A.27) implies that $|\eta(t)|_{H^{1-s}(D)} \le c\big(\frac{M}{\alpha}\big)^s h^{s^2}\ell_D^{s(1-s)}|\eta(t)|_{H^1(D)}$. Hence, we obtain

$$\|\delta_h(t)\|_{L^2(D)} \le c\,\rho^{-1}h^{s^2}\ell_D^{1-s^2}c_\kappa|\eta(t)|_{H^1(D)},$$

with $c_\kappa := \big(1+\frac{M}{\alpha}\big)\frac{M'}{\alpha} + \big(\frac{M}{\alpha}\big)^s\frac{M''}{\alpha}$. Finally, the claim follows by applying the triangle inequality to $\partial_t\eta(t) = -\delta_h(t) + (\partial_t u(t) - \Pi_h^{\mathrm{E}}(t;\partial_t u(t)))$.
(iii) Let $u \in C^\infty(\overline{J};V)$. Using the coercivity of the bilinear form a, we have

$$\begin{aligned}\alpha\,\|\partial_t\Pi_h^{\mathrm{E}}(t;u(t))\|_V^2 &\le a(t;\partial_t\Pi_h^{\mathrm{E}}(t;u(t)),\partial_t\Pi_h^{\mathrm{E}}(t;u(t)))\\ &\le a(t;\partial_t\Pi_h^{\mathrm{E}}(t;u(t)) - \Pi_h^{\mathrm{E}}(t;\partial_t u),\partial_t\Pi_h^{\mathrm{E}}(t;u(t))) + a(t;\Pi_h^{\mathrm{E}}(t;\partial_t u),\partial_t\Pi_h^{\mathrm{E}}(t;u(t))).\end{aligned}$$

Using the identity (66.19) in Lemma 66.16 and the boundedness of the bilinear forms a and $\dot{a}$, we infer that

$$\begin{aligned}\alpha\|\partial_t\Pi_h^{\mathrm{E}}(t;u(t))\|_V^2 &\le \dot{a}(t;u(t)-\Pi_h^{\mathrm{E}}(t;u(t)),\partial_t\Pi_h^{\mathrm{E}}(t;u(t)))+a(t;\Pi_h^{\mathrm{E}}(t;\partial_t u),\partial_t\Pi_h^{\mathrm{E}}(t;u(t)))\\ &\le \frac{M'}{\rho}\|u(t)-\Pi_h^{\mathrm{E}}(t;u(t))\|_V\|\partial_t\Pi_h^{\mathrm{E}}(t;u(t))\|_V+M\|\Pi_h^{\mathrm{E}}(t;\partial_t u)\|_V\|\partial_t\Pi_h^{\mathrm{E}}(t;u(t))\|_V.\end{aligned}$$

Hence, we have

$$\begin{aligned}\alpha\|\partial_t\Pi_h^{\mathrm{E}}(t;u(t))\|_V &\le \frac{M'}{\rho}\|u(t)-\Pi_h^{\mathrm{E}}(t;u(t))\|_V+M\|\Pi_h^{\mathrm{E}}(t;\partial_t u)\|_V\\ &\le \frac{M'}{\rho}\Big(1+\frac{M}{\alpha}\Big)\|u(t)\|_V+\frac{M^2}{\alpha}\|\partial_t u(t)\|_V.\end{aligned}$$

From this estimate and the Bochner theorem, we infer that

$$\partial_t\Pi_h^{\mathrm{E}}(t;u(t))\in L^2(J;V_h),$$

i.e., $\Pi_h^{\mathrm{E}}(t;u(t))\in H^1(J;V_h)$, and since $1\le\frac{M}{\alpha}$, we have

$$\|\partial_t\Pi_h^{\mathrm{E}}(\cdot;u)\|_{L^2(J;V_h)}\le\frac{M^2}{\alpha^2}\Big(\frac{2M'}{\rho M}\|u\|_{L^2(J;V)}+\|\partial_t u\|_{L^2(J;V)}\Big).$$

This implies that $\|\Pi_h^{\mathrm{E}}(t;u(t))\|_{H^1(J;V_h)}\le c(\alpha,M,\frac{M'}{\rho})\|u\|_{H^1(J;V)}$.

Chapter 67

Implicit and explicit Euler schemes

Exercises

Exercise 67.1 (Incremental Gronwall's lemma). Let $\gamma \in \mathbb{R}$, $\gamma > -1$. Let $(a_n)_{n\in\overline{\mathcal{N}}_\tau}$, $(b_n)_{n\in\mathcal{N}_\tau}$ be two sequences of real numbers s.t. $(1+\gamma)a_n \le a_{n-1} + b_n$ for all $n \in \mathcal{N}_\tau$. Prove that $a_n \le \frac{a_0}{(1+\gamma)^n} + \sum_{k\in\{1:n\}} \frac{b_k}{(1+\gamma)^{n-k+1}}$ for all $n \in \mathcal{N}_\tau$. (*Hint*: by induction.) *Note*: it is common to use the above estimate together with the inequality $\frac{1}{1+\gamma} \le e^{-\frac{\gamma}{2}}$ for $\gamma \in (0,1)$. The reader is referred to Exercise 68.3 for a discrete form of the Gronwall using an assumption that is weaker than requesting that $(1+\gamma)a_n \le a_{n-1} + b_n$.

Exercise 67.2 (Inf-sup condition). Let $X_{h\tau} := (V_h)^{N+1}$ and $Y_{h\tau} := V_h \times (V_h)^N$. Define $\|\phi_h\|_{V_h'} := \sup_{v_h\in V_h} \frac{|(\phi_h, v_h)_L|}{\|v_h\|_V}$ for all $\phi_h \in V_h$ and consider the following norms:

$$\|v_{h\tau}\|_{X_{h\tau}}^2 := \frac{1}{\alpha}\|v_h^N\|_L^2 + \|v_{h\tau}\|_{\ell^2(J;V)}^2 + \frac{1}{\alpha M}\|\delta_\tau v_{h\tau}\|_{\ell^2(J;V_h')}^2 + \frac{\tau}{\alpha}\|\delta_\tau v_{h\tau}\|_{\ell^2(J;L)}^2,$$
$$\|y_{h\tau}\|_{Y_{h\tau}}^2 := \frac{1}{\alpha}\|y_{0h}\|_L^2 + \|y_{1h\tau}\|_{\ell^2(J;V)}^2,$$

with $(\delta_\tau v_{h\tau})^n := \frac{1}{\tau}(v_h^n - v_h^{n-1})$, for all $v_{h\tau} \in X_{h\tau}$ and all $y_{h\tau} := (y_{0h}, y_{1h\tau}) \in Y_{h\tau}$. Define the bilinear form $b_\tau : X_{h\tau}\times Y_{h\tau} \to \mathbb{R}$ s.t.

$$b_\tau(v_{h\tau}, y_{h\tau}) := (v_h^0, y_{0h})_L + \sum_{n\in\mathcal{N}_\tau} \tau\big(((\delta_\tau v_{h\tau})^n, y_{1h}^n)_L + a^n(v_h^n, y_{1h}^n)\big).$$

Assume that a is symmetric. The goal is to prove the following inf-sup condition:

$$\inf_{v_{h\tau}\in X_{h\tau}} \sup_{y_h\in Y_{h\tau}} \frac{|b_\tau(v_{h\tau}, y_{h\tau})|}{\|v_{h\tau}\|_{X_{h\tau}}\|y_{h\tau}\|_{Y_{h\tau}}} \ge \alpha\left(\frac{\alpha}{M}\right)^{\frac{1}{2}}. \tag{67.1}$$

(i) Let $A_h^n : V_h \to V_h'$ be s.t. $\langle A_h^n(z_h), w_h\rangle_{V_h',V_h} := a^n(z_h, w_h)$ for all $z_h, w_h \in V_h$ and all $n \in \mathcal{N}_\tau$. Consider the test function $w_{h\tau} := (w_{0h}, w_{1h\tau}) \in Y_{h\tau}$ with $w_{0h} := v_h^0$ and $w_{1h}^n := (A_h^n)^{-1}((\delta_\tau v_{h\tau})^n) + v_h^n$ for all $n \in \mathcal{N}_\tau$. Prove that $b_\tau(v_{h\tau}, w_{h\tau}) \ge \alpha\|v_{h\tau}\|_{X_{h\tau}}^2$. (*Hint*: use that

$(A_h^n)^{-1}$ is coercive on V_h' with constant M^{-1}, see Lemma C.63.) (ii) Prove that $\alpha\tau\|w_{1h}^n\|_V^2 \le M\tau\|v_h^n\|_V^2 + \frac{\tau}{\alpha}\|(\delta_\tau v_{h\tau})^n\|_{V_h'}^2 + \|v_h^n\|_L^2 - \|v_h^{n-1}\|_L^2 + \tau^2\|(\delta_\tau v_{h\tau})^n\|_L^2$. (*Hint*: use the boundedness of $(A_h^n)^{-1}$ on V_h' with constant α^{-1}.) (iii) Conclude. *Note*: let $\mathfrak{T}_1 := \tau\|\delta_\tau u_{h\tau}\|_{\ell^2(J;L)}^2$ and consider the bound on $\mathfrak{T}_1$ given in Lemma 67.3. Let $\mathfrak{T}_2 := \frac{1}{M}\|\delta_\tau u_{h\tau}\|_{\ell^2(J;V_h')}^2$ and consider the bound on $\mathfrak{T}_2$ given by the inf-sup condition (67.1) (see Exercise 71.8). If the functions $(\partial_t u(t_n))_{n\in\mathcal{N}_\tau}$ are smooth in space for all $n \in \mathcal{N}_\tau$, one expects that $\mathfrak{T}_2 \approx \frac{\iota_{L,V}^2}{M}\|\delta_\tau u_{h\tau}\|_{\ell^2(J;L)}^2 = \frac{\rho}{2\tau}\frac{\alpha}{M}\mathfrak{T}_1$ with the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$. Hence, $\mathfrak{T}_2 \gg \mathfrak{T}_1$ if $\rho \gg \tau$, i.e., controlling $\mathfrak{T}_2$ is more informative than just controlling $\mathfrak{T}_1$.

Exercise 67.3 (Implicit-explicit scheme). Let $(V, L \equiv L', V')$ be a Gelfand triple. Let $B \in \mathcal{L}(V;L)$ and $A \in \mathcal{L}(V;V')$ be two operators. Assume that A is V-coercive with $\langle A(v), v\rangle_{V',V} \ge \alpha\|v\|_V^2$ for all $v \in V$, and that $\|v\|_L \le \iota_{L,V}\|v\|_V$. Let $\mathfrak{c}$ be s.t. $\mathfrak{c} \ge \max(\|B\|_{\mathcal{L}(V;L)}, \|B^*\|_{\mathcal{L}(L;V')})$. Let $u_0 \in V$ and $f \in C^0(\overline{J};V')$. Consider the model problem $\partial_t u(t) + A(u)(t) + B(u)(t) = f(t)$ in $L^2(J;V')$, and $u(0) = u_0$. (i) Let $\nu > 0$, $\boldsymbol{\beta} \in \boldsymbol{W}^{1,\infty}(D)$, $u_0 \in L^2(D)$, and $f \in C^0(\overline{J};H^{-1}(D))$. Show that the time-dependent advection-diffusion equation $\partial_t u - \nu\Delta u + \boldsymbol{\beta}\cdot\nabla u = f$, $u_{|\partial D} = 0$, $u(0) = u_0$ fits the above setting, i.e., specify the spaces V, L, the operators A, B, and the constants α, $\mathfrak{c}$ in this case. (ii) Let $f^n := f(t_n)$ for all $n \in \mathcal{N}_\tau$. Consider the following scheme: $u^0 := u_0$ and for all $v \in V$ and all $n \in \mathcal{N}_\tau$,

$$(u^n - u^{n-1}, v)_L + \tau\langle A(u^n), v\rangle_{V',V} + \tau(B(u^{n-1}), v)_L = \tau\langle f^n, v\rangle_{V',V}.$$

Prove that if $2\frac{\mathfrak{c}\iota_{L,V}}{\alpha} \le 1$, then

$$\|u^n\|_L^2 + \alpha\tau\|u^n\|_V^2 \le \|u^{n-1}\|_L^2 + \frac{1}{2}\alpha\tau\|u^{n-1}\|_V^2 + 2\frac{\tau}{\alpha}\|f^n\|_{V'}^2.$$

(iii) Assume that $(B(v), v)_L \ge 0$ for all $v \in V$, and that the time step satisfies the bound $\tau \le \frac{1}{2}\frac{\alpha}{\mathfrak{c}^2}$. (We no longer assume that $2\frac{\mathfrak{c}\iota_{L,V}}{\alpha} \le 1$.) Prove that

$$\|u^n\|_L^2 + \alpha\tau\|u^n\|_V^2 \le \|u^{n-1}\|_L^2 + \frac{1}{2}\alpha\tau\|u^{n-1}\|_V^2 + \frac{\tau}{\alpha}\|f^n\|_{V'}^2.$$

Solution to exercises

Exercise 67.1 (Incremental Gronwall's lemma). We proceed by induction. For $n = 1$, we have $a_1 \le \frac{a_0}{(1+\gamma)^1} + \frac{b_1}{(1+\gamma)}$ which is exactly the formula that we want to prove. Let us assume that $a_n \le \frac{a_0}{(1+\gamma)^n} + \sum_{k\in\{1:n\}}\frac{b_k}{(1+\gamma)^{n-k+1}}$, for some $n \in \mathcal{N}_\tau$, $n < N$. We have

$$\begin{aligned} a_{n+1} &\le \frac{a_n}{1+\gamma} + \frac{b_{n+1}}{1+\gamma} \\ &\le \frac{a_0}{(1+\gamma)^n(1+\gamma)} + \frac{b_{n+1}}{1+\gamma} + \sum_{k\in\{1:n\}}\frac{b_k}{(1+\gamma)^{n-k+2}} \\ &\le \frac{a_0}{(1+\gamma)^{n+1}} + \sum_{k\in\{1:n+1\}}\frac{b_k}{(1+\gamma)^{n-k+2}}. \end{aligned}$$

Exercise 67.2 (Inf-sup condition). Let $v_{h\tau} \in X_{h\tau}$ and let us set $w_{h\tau} := (w_{0h}, w_{1h\tau}) \in Y_{h\tau}$ with $w_{0h} := v_h^0$ and $w_{1h}^n := (A_h^n)^{-1}((\delta_\tau v_{h\tau})^n) + v_h^n$ for all $n \in \mathcal{N}_\tau$. Notice that A_h^n is self-adjoint

by assumption.
(i) A straightforward calculation shows that

$$\begin{aligned} b_\tau(v_{h\tau},w_{h\tau}) &= \|v_h^0\|_L^2 + \sum_{n\in\mathcal{N}_\tau}\Big((v_h^n - v_h^{n-1}, w_{1h}^n)_L + \tau a^n(v_h^n, w_{1h}^n)\Big) \\ &= \|v_h^0\|_L^2 + \sum_{n\in\mathcal{N}_\tau}\Big(2(v_h^n - v_h^{n-1}, v_h^n)_L + \tau\langle A_h^n(v_h^n), v_h^n\rangle_{V_h',V_h} \\ &\quad + \tau^{-1}\langle v_h^n - v_h^{n-1}, (A_h^n)^{-1}(v_h^n - v_h^{n-1})\rangle_{V_h',V_h}\Big). \end{aligned}$$

Using the coercivity of A_h^n on V_h (with constant α) and the coercivity of $(A_h^n)^{-1}$ on V_h' (with constant M^{-1} owing to Lemma C.63), and using the identity (67.9) yields

$$\begin{aligned} b_\tau(v_{h\tau},w_{h\tau}) &\ge \|v_h^0\|_L^2 + \sum_{n\in\mathcal{N}_\tau}\Big(\|v_h^n\|_L^2 - \|v_h^{n-1}\|_L^2 + \|v_h^n - v_h^{n-1}\|_L^2 \\ &\quad + \alpha\tau\|v_h^n\|_V^2 + \frac{1}{M\tau}\|v_h^n - v_h^{n-1}\|_{V_h'}^2\Big) \ge \alpha\|v_{h\tau}\|_{X_{h\tau}}^2. \end{aligned}$$

(ii) Another straightforward computation using the coercivity of A_h^n, its boundedness on V_h (with constant M), the boundedness of $(A_h^n)^{-1}$ on V_h' (with constant α^{-1}), and the self-adjointness of A_h^n shows that

$$\begin{aligned} \alpha\tau\|w_{1h}^n\|_V^2 &\le \tau\langle A_h^n(w_{1h}^n), w_{1h}^n\rangle_{V_h',V_h} \\ &\le \tau^{-1}\langle v_h^n - v_h^{n-1} + \tau A_h^n(v_h^n), (A_h^n)^{-1}(v_h^n - v_h^{n-1}) + \tau v_h^n\rangle_{V_h',V_h} \\ &\le M\tau\|v_h^n\|_V^2 + \frac{1}{\alpha\tau}\|v_h^n - v_h^{n-1}\|_{V_h'}^2 + \|v_h^n\|_L^2 - \|v_h^{n-1}\|_L^2 + \|v_h^n - v_h^{n-1}\|_L^2 \\ &= M\tau\|v_h^n\|_V^2 + \frac{\tau}{\alpha}\|(\delta_\tau v_{h\tau})^n\|_{V_h'}^2 + \|v_h^n\|_L^2 - \|v_h^{n-1}\|_L^2 + \tau^2\|(\delta_\tau v_{h\tau})^n\|_L^2. \end{aligned}$$

(iii) Summing the estimate from Step (ii) over $n\in\mathcal{N}_\tau$, we obtain

$$\begin{aligned} \alpha\|w_{h\tau}\|_{Y_{h\tau}}^2 &= \|w_{0h}\|_L^2 + \sum_{n\in\mathcal{N}_\tau}\alpha\tau\|w_{1h}^n\|_V^2 \\ &\le M\|v_{h\tau}\|_{\ell^2(J;V)}^2 + \frac{1}{\alpha}\|\partial_\tau v_{h\tau}\|_{\ell^2(J;V_h')}^2 + \tau\|\partial_\tau v_{h\tau}\|_{\ell^2(J;L)}^2 + \|v_h^N\|_L^2. \end{aligned}$$

Hence, $\frac{\alpha}{M}\|w_{h\tau}\|_{Y_{h\tau}}^2 \le \|v_{h\tau}\|_{X_{h\tau}}^2$ since $\alpha\le M$, and the assertion follows.

Exercise 67.3 (Implicit-explicit scheme). (i) Let $\nu>0$, $\boldsymbol{\beta}\in\boldsymbol{W}^{1,\infty}(D)$. The time-dependent advection-diffusion equation $\partial_t u - \nu\Delta u + \boldsymbol{\beta}\cdot\nabla u = f$ for a.e. $(\boldsymbol{x},t)\in D\times J$, fits the proposed framework with $L:=L^2(D)$, $V:=H_0^1(D)$. The operator $A:H_0^1(D)\to H^{-1}(D)$ is s.t. $A(v):=-\nu\Delta v$, and the operator B is s.t. $B(v):=\boldsymbol{\beta}\cdot\nabla v$. Let us equip V with the H^1-seminorm, i.e., $\|v\|_V:=\|\nabla v\|_{\boldsymbol{L}^2(D)}$. Then the coercivity constant of A is $\alpha:=\nu$. Moreover, we have

$$\begin{aligned} \|B(v)\|_{V'} &= \sup_{w\in V}\frac{\langle\boldsymbol{\beta}\cdot\nabla v, w\rangle_{V',V}}{\|w\|_V} = \sup_{w\in V}\frac{-(v,\boldsymbol{\beta}\cdot\nabla w)_L - ((\nabla\cdot\boldsymbol{\beta})v, w)_L}{\|w\|_V} \\ &\le (\|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)} + \iota_{L,V}\|\nabla\cdot\boldsymbol{\beta}\|_{L^\infty(D)})\|v\|_L, \end{aligned}$$

i.e., $\|B^*\|_{\mathcal{L}(L;V')}\le\|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)} + \iota_{L,V}\|\nabla\cdot\boldsymbol{\beta}\|_{L^\infty(D)}$. Moreover, we have $\|B(v)\|_L\le\|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}\|v\|_V$, i.e., $\|B\|_{\mathcal{L}(V;L)}\le\|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$. Hence, we can take $\mathfrak{c}:=\|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)} + \iota_{L,V}\|\nabla\cdot\boldsymbol{\beta}\|_{L^\infty(D)}$.
(ii) Let us test the discrete equation with $2\tau u^n$. We obtain

$$\|u^n\|_L^2 + \|u^n - u^{n-1}\|_L^2 - \|u^{n-1}\|_L^2 + 2\alpha\tau\|u^n\|_V^2 + 2\tau(B(u^{n-1}),u^n)_L \le 2\tau\|f^n\|_{V'}\|u^n\|_V,$$

where we used the V-coercivity of A. As a result, we infer that

$$\begin{aligned}\|u^n\|_L^2 + 2\alpha\tau\|u^n\|_V^2 &\le \|u^{n-1}\|_L^2 + 2\tau\|f^n\|_{V'}\|u^n\|_V + 2\tau\|u^{n-1}\|_V\|B^*(u^n)\|_{V'}\\ &\le \|u^{n-1}\|_L^2 + 2\frac{\tau}{\alpha}\|f^n\|_{V'}^2 + \frac{1}{2}\alpha\tau\|u^n\|_V^2 + 2\tau\mathfrak{c}\|u^{n-1}\|_V\|u^n\|_L\\ &\le \|u^{n-1}\|_L^2 + 2\frac{\tau}{\alpha}\|f^n\|_{V'}^2 + \frac{1}{2}\alpha\tau\|u^n\|_V^2 + 2\tau\mathfrak{c}\iota_{L,V}\|u^{n-1}\|_V\|u^n\|_V\\ &\le \|u^{n-1}\|_L^2 + 2\frac{\tau}{\alpha}\|f^n\|_{V'}^2 + \frac{1}{2}\alpha\tau\|u^n\|_V^2 + 2\tau\frac{\mathfrak{c}^2\iota_{L,V}^2}{\alpha}\|u^{n-1}\|_V^2 + \frac{1}{2}\alpha\tau\|u^n\|_V^2.\end{aligned}$$

Rearranging the terms, we obtain

$$\|u^n\|_L^2 + \alpha\tau\|u^n\|_V^2 \le \|u^{n-1}\|_L^2 + 2\tau\frac{\mathfrak{c}^2\iota_{L,V}}{\alpha}\|u^{n-1}\|_V^2 + 2\frac{\tau}{\alpha}\|f^n\|_{V'}^2.$$

Owing to the assumption $2\frac{\mathfrak{c}^2\iota_{L,V}^2}{\alpha} \le \frac{1}{2}\alpha$, we infer that

$$\|u^n\|_L^2 + \alpha\tau\|u^n\|_V^2 \le \|u^{n-1}\|_L^2 + \frac{1}{2}\alpha\tau\|u^{n-1}\|_V^2 + 2\frac{\tau}{\alpha}\|f^n\|_{V'}^2,$$

which is the expected inequality.
(iii) Let us now assume that $(B(v), v)_L \ge 0$ for all $v \in V$. Testing again the discrete equation with $2\tau u^n$, we obtain

$$\begin{aligned}\|u^n\|_L^2 + \|u^n - u^{n-1}\|_L^2 - \|u^{n-1}\|_L^2 + 2\alpha\tau\|u^n\|_V^2 + 2\tau(B(u^{n-1}), u^n - u^{n-1})_L&\\ + 2\tau(B(u^{n-1}), u^{n-1})_L \le 2\tau\|f^n\|_{V'}\|u^n\|_V.&\end{aligned}$$

Since $(B(u^{n-1}), u^{n-1})_L \ge 0$ by assumption, rearranging the terms, and applying Young's inequality to the right-hand side, we infer that

$$\begin{aligned}&\|u^n\|_L^2 + \|u^n - u^{n-1}\|_L^2 - \|u^{n-1}\|_L^2 + 2\alpha\tau\|u^n\|_V^2\\ &\le 2\tau\|f^n\|_{V'}\|u^n\|_V - 2\tau(B(u^{n-1}), u^n - u^{n-1})_L\\ &\le \frac{\tau}{\alpha}\|f^n\|_{V'}^2 + \alpha\tau\|u^n\|_V^2 + \tau^2\|B(u^{n-1})\|_L^2 + \|u^n - u^{n-1}\|_L^2.\end{aligned}$$

After simplification, we obtain

$$\|u^n\|_L^2 + \alpha\tau\|u^n\|_V^2 \le \|u^{n-1}\|_L^2 + \tau^2\|B(u^{n-1})\|_L^2 + \frac{\tau}{\alpha}\|f^n\|_{V'}^2.$$

The expected estimate follows by using the boundedness of B and the assumption $\tau\mathfrak{c}^2 \le \frac{1}{2}\alpha$.

Chapter 68

BDF2 and Crank–Nicolson schemes

Exercises

Exercise 68.1 (Heat equation). Write the error estimates for the heat equation using the BDF2 time discretization in the setting of Remark 68.8.

Exercise 68.2 (Inverse inequality on A_h). Prove (68.22). (*Hint*: observe that $\|A_h(v_h)\|_L = \max_{w_h \in V_h} \frac{|(A_h(v_h), w_h)_L|}{\|w_h\|_L}$ and use the boundedness of a.)

Exercise 68.3 (Discrete Gronwall's lemma). The objective of this exercise is to prove the following discrete Gronwall's lemma. Let $(\gamma_n)_{n\in\mathcal{N}_\tau}$, $(a_n)_{n\in\mathcal{N}_\tau}$, $(b_n)_{n\in\mathcal{N}_\tau}$, $(c_n)_{n\in\mathcal{N}_\tau}$ be sequences of real numbers. Let $B \in \mathbb{R}$. Assume that

$$\gamma_n \in (0,1), \qquad a_n \geq 0, \qquad b_n \geq 0, \tag{68.1a}$$

$$a_n + \sum_{l\in\{1:n\}} b_l \leq \sum_{l\in\{1:n\}} \gamma_l a_l + \sum_{l\in\{1:n\}} c_l + B, \tag{68.1b}$$

for all $n \in \mathcal{N}_\tau$. Then we have

$$a_n + \sum_{l\in\{1:n\}} b_l \leq \sum_{l\in\{1:n\}} c_l \prod_{\mu\in\{l:n\}} \frac{1}{1-\gamma_\mu} + B \prod_{\mu\in\{1:n\}} \frac{1}{1-\gamma_\mu}. \tag{68.2}$$

(i) Let $d_n := \sum_{l\in\{1:n\}} \gamma_l a_l + \sum_{l\in\{1:n\}} (c_l - b_l) + B - a_n$ and let $S_n := d_n + a_n + \sum_{l\in\{1:n\}} b_l$. Show that $S_n(1-\gamma_n) \leq S_{n-1} + c_n$ for all $n \geq 2$. (*Hint*: observe that $a_n \leq S_n$.) (ii) Show by induction that $S_n \leq \sum_{l\in\{1:n\}} c_l \prod_{\mu\in\{l:n\}} \frac{1}{1-\gamma_\mu} + B \prod_{\mu\in\{1:n\}} \frac{1}{1-\gamma_\mu}$. Conclude. (*Hint*: (68.1b) means that $d_n \geq 0$.) *Note*: if one replaces the assumption (68.1b) by the assumption $(1+\gamma)a_n \leq a_{n-1} + c_n$ which implies (68.1b) with $b_l := 0$, $B := a_0$, and $\gamma_l := -\gamma$ for all $l \in \{1{:}n\}$, the incremental Gronwall lemma from Exercise 67.1 leads to the same bound on a_n as (68.2). The incremental Gronwall lemma only requires that $\gamma > -1$, whereas the discrete Gronwall lemma requires that $\gamma_l \in (0,1)$ (i.e., $\gamma \in (-1,0)$ if one sets $\gamma_l := \gamma$).

Exercise 68.4 (Variant on BDF2). The objective of this exercise is to revisit the stability argument for BDF2 proposed in Thomée [43, p. 18]. Consider the setting introduced in §68.2 and the scheme (68.1). (i) Show that for all $k \geq 2$

$$(\tfrac{3}{2}u_h^k - 2u_h^{k-1} + \tfrac{1}{2}u_h^{k-2}, u_h^k)_L = \|u_h^k\|_L^2 - \|u_h^{k-1}\|_L^2 - \tfrac{1}{4}(\|u_h^k\|_L^2 - \|u_h^{k-2}\|_L^2) + \|u_h^k - u_h^{k-1}\|_L^2 - \tfrac{1}{4}\|u_h^k - u_h^{k-2}\|_L^2.$$

(ii) Prove that $\sum_{k\in\{2:n\}} \|u_h^k\|_L^2 - \|u_h^{k-1}\|_L^2 - \frac{1}{4}(\|u_h^k\|_L^2 - \|u_h^{k-2}\|_L^2) = \frac{3}{4}\|u_h^n\|_L^2 - \frac{1}{4}\|u_h^{n-1}\|_L^2 - \frac{3}{4}\|u_h^1\|_L^2 + \frac{1}{4}\|u_h^0\|_L^2$, and that

$$\sum_{k\in\{2:n\}} \|u_h^k - u_h^{k-1}\|_L^2 - \tfrac{1}{4}\|u_h^k - u_h^{k-2}\|_L^2 \geq \tfrac{1}{2}\|u_h^n - u_h^{n-1}\|_L^2 - \tfrac{1}{2}\|u_h^1 - u_h^0\|_L^2.$$

(iii) Show that

$$(u_h^1 - u_h^0, u_h^1)_L + \sum_{k\in\{2:n\}} (\tfrac{3}{2}u_h^k - 2u_h^{k-1} + \tfrac{1}{2}u_h^{k-2}, u_h^k)_L \geq \tfrac{3}{4}\|u_h^n\|_L^2 - \tfrac{1}{4}\|u_h^{n-1}\|_L^2 - \tfrac{1}{4}\|u_h^1\|_L^2 - \tfrac{1}{4}\|u_h^0\|_L^2.$$

(iv) Assuming that $f^k \in L$ for all $k \in \mathcal{N}_\tau$, show that

$$3\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 + \sum_{k\in\{1:n\}} 4\tau\alpha\|u_h^k\|_V^2 \leq \|u_h^0\|_L^2 + \|u_h^1\|_L^2 + \sum_{k\in\{1:n\}} 4\tau\|f^k\|_L\|u_h^k\|_L.$$

(v) Letting $m \in \{0{:}n\}$ be the index s.t. $\|u_h^m\|_L := \|u_{h\tau}\|_{\ell^\infty(\overline{J};L)}$, show that

$$2\|u_{h\tau}\|_{\ell^\infty(\overline{J};L)} \leq \|u_h^0\|_L + \|u_h^1\|_L + \sum_{k\in\{1:n\}} 4\tau\|f^k\|_L.$$

(vi) Conclude that $\|u_{h\tau}\|_{\ell^\infty(\overline{J};L)} \leq \|u_h^0\|_L + \frac{\tau}{2}\|f^1\|_L + \sum_{k\in\{1:n\}} 2\tau\|f^k\|_L$.
(vii) Modify the argument to account for $f^k \in V'$ instead of $f^k \in L$ for all $k \geq 2$, and $f^1 = f_{V'}^1 + f_L^1$, where $f_{V'}^1 \in V'$ and $f_L^1 \in L$, and prove that

$$\|u_{h\tau}\|_{\ell^\infty(\overline{J};L^2)}^2 \leq \frac{5}{2}\|u_h^0\|_L^2 + 6\tau^2\|f_L^1\|_L^2 + \sum_{k\in\{1:n\}} \frac{\tau}{\alpha}\|\tilde{f}^k\|_{V'}^2.$$

Exercise 68.5 (Variant of Crank–Nicolson scheme). Consider the following variant of the Crank–Nicolson scheme: after setting $u_h^0 := \mathcal{P}_{V_h}(u^0)$, we construct the sequence of functions $u_{h\tau} := (u_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$ such that

$$(u_h^n - u_h^{n-1}, w_h)_L + \frac{1}{2}\tau\big(a^n(u_h^n, w_h) + a^{n-1}(u_h^{n-1}, w_h)\big) = \tau\langle \check{f}^{n-\frac{1}{2}}, w_h\rangle_{V',V},$$

for all $w_h \in V_h$ and all $n \in \mathcal{N}_\tau$, with $a^n(\cdot,\cdot) := a(t_n;\cdot,\cdot)$, $a^{n-1}(\cdot,\cdot) := a(t_{n-1};\cdot,\cdot)$, and $\check{f}^{n-\frac{1}{2}} := \frac{1}{2}(f(t_n) + f(t_{n-1})) \in V'$. Assume that $f \in C^0(\overline{J};L)$ and that the restriction (68.20) on the time step holds true. Prove again the bound (68.21) on $\|u_h^n\|_L^2$ with $\check{f}^{k-\frac{1}{2}}$ in lieu of $f^{k-\frac{1}{2}}$ on the right-hand side. (*Hint*: adapt the proof of Lemma 68.12 by starting from the identity $u_h^n + \frac{1}{2}\tau A_h^n(u_h^n) = u_h^{n-1} - \frac{1}{2}\tau A_h^{n-1}(u_h^{n-1}) + \check{f}^{n-\frac{1}{2}}$.) *Note*: deriving an $\ell^2(J;V)$-stability estimate as in Lemma 68.9 is more delicate with this variant of the Crank–Nicolson scheme.

Solution to exercises

Exercise 68.1 (Heat equation). Using the estimate (68.10) from Theorem 68.4 and the approximation properties of the finite element space V_h, we infer that there is c s.t. for all $h \in \mathcal{H}$, τ, α, and M,

$$\begin{aligned}\|u_\tau - u_{h\tau}\|_{\ell^2(J;V)} \le c\,\bigg(&\frac{1}{\sqrt{\alpha}}\tau^2\|\partial_{tt}u\|_{C^0(\overline{J}_1;L)} + \frac{\rho}{\iota_{L,V}}\tau^2\|\partial_{ttt}u\|_{L^2(J;L)}\\ &+ h^r\Big(\frac{1}{\sqrt{\alpha}}\|u_0\|_{H^r(D)} + \frac{M}{\alpha}|u_\tau|_{\ell^2(J;H^{r+1}(D))}\\ &+ \frac{\rho}{\iota_{L,V}}|\partial_t u|_{L^2(J;H^r(D))}\Big)\bigg).\end{aligned}$$

Notice that the error estimates in space can be localized over the mesh cells. If in addition the bilinear form a is time-independent and $\tau \le \frac{1}{\rho}$, the estimate (68.15) from Theorem 68.7 gives

$$\begin{aligned}\|u_h^n - u(t_n)\|_L \le c\,h^{r+s}\bigg(&\|u(t_n)\|_{H^{r+1}(D)}\\ &+ c_1\Big(e^{-\frac{t_n}{8\rho}}\|u_0\|_{H^{r+1}(D)} + \sqrt{\rho}\|e^{-\frac{t_n-\cdot}{8\rho}}\partial_t u\|_{L^2((0,t_n);H^{r+1}(D))}\Big)\bigg)\\ &+ c_2\,\tau^2\Big(e^{-\frac{t_n}{8\rho}}\|\partial_{tt}u\|_{C^0(\overline{J}_1;L)} + \sqrt{\rho}\|e^{-\frac{t_n-\cdot}{8\rho}}\partial_{ttt}u\|_{L^2((0,t_n);L)}\Big).\end{aligned}$$

Exercise 68.2 (Inverse inequality). Lt $v_h \in V_h$. Using the definition of A_h, we have

$$\begin{aligned}\|A_h(v_h)\|_L &= \max_{w_h\in V_h}\frac{|(A_h(v_h),w_h)_L|}{\|w_h\|_L} = \max_{w_h\in V_h}\frac{|a(v_h,w_h)|}{\|w_h\|_L}\\ &\le M\|v_h\|_V\max_{w_h\in V_h}\frac{\|w_h\|_V}{\|w_h\|_L} = \iota_{L,V}^{-1}c_{\mathrm{INV}}(h)M\|v_h\|_V,\end{aligned}$$

whence the assertion.

Exercise 68.3 (Discrete Gronwall's lemma). (i) Using the definition of S_n, we have $S_n = B + \sum_{l\in\{1:n\}}\gamma_l a_l + \sum_{l\in\{1:n\}} c_l$, i.e.,

$$S_n - S_{n-1} = \gamma_n a_n + c_n.$$

But $S_n = d_n + a_n + \sum_{l\in\{1:n\}} b_l \ge a_n$, since $d_n \ge 0$ and $b_l \ge 0$ by the assumptions (68.1b) and (68.1a), respectively. Using that $0 \le \gamma_n$ and $0 \le a_n \le S_n$, owing to (68.1a), we infer that

$$S_n(1-\gamma_n) \le S_{n-1} + c_n.$$

(ii) For $s = 1$, we have

$$S_1 = a_1 + b_1 + d_1 = \gamma_1 a_1 + c_1 + B \le \gamma_1 S_1 + c_1 + B,$$

since $\gamma_1 \ge 0$. Hence, $S_1 \le \frac{c_1}{1-\gamma_1} + B\frac{1}{1-\gamma_1}$, which is the expected result. Assume now that $n \ge 2$ and that the induction assumption holds true for $n-1$. The inequality $S_n(1-\gamma_n) \le S_{n-1} + c_n$ implies that

$$\begin{aligned}S_n &\le \frac{c_n}{1-\gamma_n} + \sum_{l\in\{1:n-1\}} c_l\frac{1}{1-\gamma_n}\prod_{\mu\in\{l:n-1\}}\frac{1}{1-\gamma_\mu} + B\frac{1}{1-\gamma_n}\prod_{\mu\in\{1:n-1\}}\frac{1}{1-\gamma_\mu}\\ &\le \sum_{l\in\{1:n\}} c_l\prod_{\mu\in\{l:n\}}\frac{1}{1-\gamma_\mu} + B\prod_{\mu\in\{1:n\}}\frac{1}{1-\gamma_\mu},\end{aligned}$$

thereby proving that $S_n \le \sum_{l\in\{1:n\}} c_l \prod_{\mu\in\{l:n\}} \frac{1}{1-\gamma_\mu} + B\prod_{\mu\in\{1:n\}} \frac{1}{1-\gamma_\mu}$ for all $n \in \mathcal{N}_\tau$. This estimate, in turn, proves (68.2) since $S_n \ge a_n + \sum_{l\in\{1:n\}} b_l$ because the assumption (68.1b) is equivalent to $d_n \ge 0$ for all $n \in \mathcal{N}_\tau$.

Exercise 68.4 (Variant on BDF2). (i) We write

$$\frac{3}{2}u_h^k - 2u_h^{k-1} + \frac{1}{2}u_h^{k-2} = 2(u_h^k - u_h^{k-1}) - \frac{1}{2}(u_h^k - u_h^{k-2}).$$

Then we use the identity $(a-b,a)_L = \frac{1}{2}\|a\|_L^2 + \frac{1}{2}\|a-b\|_L^2 - \frac{1}{2}\|b\|_L^2$ to obtain the expected result.
(ii) The first identity is just a telescoping sum. For the second inequality, we use the Cauchy–Schwarz and Young's inequalities as follows:

$$\begin{aligned}-\frac{1}{4}\|u_h^k - u_h^{k-2}\|_L^2 &= -\frac{1}{4}\|u_h^k - u_h^{k-1}\|_L^2 - \frac{1}{2}(u_h^k - u_h^{k-1}, u_h^{k-1} - u_h^{k-2})_L \\ &\quad - \frac{1}{4}\|u_h^{k-1} - u_h^{k-2}\|_L^2 \ge -\frac{1}{2}\|u_h^k - u_h^{k-1}\|_L^2 - \frac{1}{2}\|u_h^{k-1} - u_h^{k-2}\|_L^2.\end{aligned}$$

The telescoping sum argument leads to the expected result.
(iii) Using $(u_h^1 - u_h^0, u_h^1)_L = \frac{1}{2}\|u_h^1\|_L^2 + \frac{1}{2}\|u_h^1 - u_h^0\|_L^2 - \frac{1}{2}\|u_h^0\|_L^2$ together with the two identities established in Step (ii), we obtain

$$\begin{aligned}&(u_h^1 - u_h^0, u_h^1)_L + \sum_{k\in\{2:n\}} (\tfrac{3}{2}u_h^k - 2u_h^{k-1} + \tfrac{1}{2}u_h^{k-2}, u_h^k)_L \\ &\quad \ge \tfrac{1}{2}\|u_h^1\|_L^2 + \tfrac{1}{2}\|u_h^1 - u_h^0\|_L^2 - \tfrac{1}{2}\|u_h^0\|_L^2 + \tfrac{3}{4}\|u_h^n\|_L^2 - \tfrac{1}{4}\|u_h^{n-1}\|_L^2 \\ &\qquad - \tfrac{3}{4}\|u_h^1\|_L^2 + \tfrac{1}{4}\|u_h^0\|_L^2 + \tfrac{1}{2}\|u_h^n - u_h^{n-1}\|_L^2 - \tfrac{1}{2}\|u_h^1 - u_h^0\|_L^2 \\ &\quad \ge \tfrac{3}{4}\|u_h^n\|_L^2 - \tfrac{1}{4}\|u_h^{n-1}\|_L^2 - \tfrac{1}{4}\|u_h^1\|_L^2 - \tfrac{1}{4}\|u_h^0\|_L^2.\end{aligned}$$

(iv) Using u_h^1 and u_h^n as the test functions in (68.1) together with the coercivity of a^n and the Cauchy–Schwarz inequality, the lower bound from Step (iii) gives

$$3\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 + \sum_{k\in\{1:n\}} 4\tau\alpha\|u_h^k\|_V^2 \le \|u_h^0\|_L^2 + \|u_h^1\|_L^2 + \sum_{k\in\{1:n\}} 4\tau\|f^k\|_L\|u_h^k\|_L.$$

(v) Letting $m \in \{1{:}n\}$ be the index s.t. $\|u_h^m\|_L := \|u_{h\tau}\|_{\ell^\infty(\overline{J};L)}$, we obtain

$$3\|u_h^m\|_L^2 \le \|u_h^m\|_L^2 + (\|u_h^1\|_L + \|u_h^0\|_L + \sum_{k\in\{1:n\}} 4\tau\|f^k\|_L)\|u_h^m\|_L.$$

This proves the expected bound.
(vi) Using the estimate $\|u_h^1\|_L^2 \le \|u_h^0\|_L\|u_h^1\|_L + \tau\|f^1\|_L\|u_h^1\|_L$ (which follows by using again the test function u_h^1 in the first implicit Euler step and the coercivity of a^1), we infer that $\|u_h^1\|_L \le \|u_h^0\|_L + \tau\|f^1\|_L$. Combined with the estimate from Step (v), this leads to the expected bound.
(vii) To account for $f^k \in V'$ (instead of $f^k \in L$), for all $k \ge 2$, and $f^1 := f_{V'}^1 + f_L^1$ with $f_{V'}^1 \in V'$ and $f_L^1 \in L$, we modify the argument from Step (iv) as follows:

$$\begin{aligned}3\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 + \sum_{k\in\{1:n\}} 4\tau\alpha\|u_h^k\|_V^2 &\le \|u_h^0\|_L^2 + \|u_h^1\|_L^2 \\ &\quad + 4\tau\|f_L^1\|_L\|u_h^1\|_L + \sum_{k\in\{1:n\}} 4\tau\|\tilde{f}^k\|_{V'}\|u_h^k\|_V.\end{aligned}$$

Using the inequalities $4\tau\|f_L^1\|_L\|u_h^1\|_L \le \|u_h^1\|_L^2 + 4\tau^2\|f_L^1\|_L^2$ together with $4\tau\|\tilde f^k\|_{V'}\|u_h^k\|_V \le 4\tau\alpha\|u_h^k\|_V^2 + \frac{\tau}{\alpha}\|\tilde f^k\|_{V'}^2$ for all $k \in \mathcal{N}_\tau$ gives

$$3\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 \le \|u_h^0\|_L^2 + 2\|u_h^1\|_L^2 + 4\tau^2\|f_L^1\|_L^2 + \sum_{k\in\{1:n\}} \frac{\tau}{\alpha}\|\tilde f^k\|_{V'}^2.$$

Moreover, using the test function u_h^1 in the first implicit Euler step, the coercivity of a^1 and the Cauchy–Schwarz inequality leads to

$$\frac12\|u_h^1\|_L^2 - \frac12\|u_h^0\|_L^2 + \frac12\|u_h^1 - u_h^0\|_L^2 + \tau\alpha\|u_h^1\|_V^2 \le \tau\|f_L^1\|_L\|u_h^1\|_L + \tau\|f_{V'}^1\|_{V'}\|u_h^1\|_V.$$

Using Young's inequalities on the right-hand side and discarding the nonnegative term $\frac12\|u_h^1-u_h^0\|_L^2$ from the left-hand side, we infer that

$$\frac12\|u_h^1\|_L^2 + \tau\alpha\|u_h^1\|_V^2 \le \frac12\|u_h^0\|_L^2 + \frac14\|u_h^1\|_L^2 + \tau^2\|f^1\|_L^2 + \tau\alpha\|u_h^1\|_V^2 + \frac{\tau}{4\alpha}\|\tilde f^1\|_{V'}^2,$$

which gives $\|u_h^1\|_L^2 \le 2\|u_h^0\|_L^2 + 4\tau^2\|f^1\|_L^2 + \frac{\tau}{\alpha}\|\tilde f^1\|_{V'}^2$. Hence, we have

$$3\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 \le 5\|u_h^0\|_L^2 + 12\tau^2\|f_L^1\|_L^2 + \sum_{k\in\{1:n\}} \frac{2\tau}{\alpha}\|\tilde f^k\|_{V'}^2.$$

Using the same argument as in Step (v) leads to the expected bound.

Exercise 68.5 (Variant of Crank–Nicolson scheme). For all $n \in \mathcal{N}_\tau$, let us define the linear operator $A_h^n : V_h \to V_h$ by setting $(A_h^n(v_h), w_h)_L := a(t_n; v_h, w_h)$ for all $v_h, w_h \in V_h$, and let us set $\check f_h^{n-\frac12} := \mathcal{P}_{V_h}(\check f^{n-\frac12})$. Then the modified Crank–Nicolson scheme can be rewritten as

$$u_h^n + \frac12\tau A_h^n(u_h^n) = u_h^{n-1} - \frac12\tau A_h^{n-1}(u_h^{n-1}) + \check f_h^{n-\frac12}.$$

Squaring this equality, we obtain

$$\begin{aligned}
&\|u_h^n\|_L^2 + \tau a^n(u_h^n, u_h^n) + \frac14\tau^2\|A_h^n(u_h^n)\|_L^2\\
&\quad= \|u_h^{n-1}\|_L^2 - \tau a^{n-1}(u_h^{n-1}, u_h^{n-1}) + \frac14\tau^2\|A_h^{n-1}(u_h^{n-1})\|_L^2\\
&\qquad+ 2\tau(u_h^{n-1}, \check f_h^{n-\frac12})_L - \tau^2(A_h^{n-1}(u_h^{n-1}), \check f_h^{n-\frac12})_L + \tau^2\|\check f_h^{n-\frac12}\|_L^2.
\end{aligned}$$

Proceeding as in the proof of Lemma 68.12, we infer that

$$\|u_h^n\|_L^2 + \alpha\tau\|u_h^n\|_V^2 + \frac{\tau^2}{4}\|A_h^n(u_h^n)\|_L^2 \le \|u_h^{n-1}\|_L^2 + \frac{\frac{\tau^2}{4}}{1+\frac{2\tau}{\rho}}\|A_h^{n-1}(u_h^{n-1})\|_L^2 + \frac72\tau\rho\|\check f_h^{n-\frac12}\|_L^2.$$

This estimate takes again the form $(1+\gamma)a^n \le a^{n-1} + b^n$ with $a^n := \|u_h^n\|_L^2 + \frac{\tau^2}{4(1+\gamma)}\|A_h^n(u_h^n)\|_L^2$ and $b^n := \frac72\tau\rho\|\check f_h^{n-\frac12}\|_L^2$. We can now conclude as before.

Chapter 69

Discontinuous Galerkin in time

Exercises

Exercise 69.1 (Integral identities). Prove the identities (69.11). (*Hint*: use that the Gauss–Radau quadrature is of order $2k$.)

Exercise 69.2 (Equivalence with Radau IIA IRK). Prove the converse assertion in Lemma 69.11. (*Hint*: show that

$$\mathcal{R}_\tau(u_{h\tau})(t) = u_h^{n-1} + \tau \sum_{j\in\{1:k+1\}} \frac{1}{2}\int_{-1}^{T_n^{-1}(t)} \mathcal{L}_j(\xi)\,\mathrm{d}\xi\big(f_h(t_{n,j}) - \mathcal{A}_h(t_{n,j})(u_h^{n,j})\big),$$

for all $t \in J_n$.)

Exercise 69.3 (Poincaré in time). Let $n \in \mathcal{N}_\tau$ and H be a Hilbert space. Show that $\|v\|^2_{L^2(J_n;H)} \le 2\tau\|v(t_{n-1}^+)\|_H^2 + \tau^2\|\partial_t v\|^2_{L^2(J_n;H)}$ for all $v \in H^1(J_n;H)$. (*Hint*: use that $v(t) = v(t_{n-1}^+) + \int_{t_{n-1}}^t \partial_t v \mathrm{d}t$ for all $t \in J_n$.)

Exercise 69.4 (Time reconstruction). (i) Show that the definition of R_τ given in Remark 69.9 is equivalent to Definition 69.5. (ii) Show that the two definitions of θ_{k+1} given in Remark 69.9 are identical. (*Hint*: set $\delta(s) := \frac{(-1)^k}{2}(L_k - L_{k+1}) - \prod_{l\in\{1:k+1\}} \frac{\xi_l - s}{\xi_l + 1}$ and prove that $\delta(-1) = 0$ and $\int_{\widehat{J}} \delta'(s)q(s)\,\mathrm{d}s = 0$ for all $q \in \mathbb{P}_k(\widehat{J};\mathbb{R})$.) (iii) Let $(V, L \equiv L', V')$ be a Gelfand triple. Let $\widehat{\mathcal{R}} : \mathbb{P}_k(\widehat{J};\mathbb{R}) \to \mathbb{P}_{k+1}(\widehat{J};\mathbb{R})$ be s.t. $\widehat{\mathcal{R}}(q) := q - q(-1)\theta_{k+1}$. Let $\mathcal{R}_n : \mathbb{P}_k(J_n;\mathbb{R}) \to \mathbb{P}_{k+1}(J_n;\mathbb{R})$ be s.t. $\mathcal{R}_n(v) = \sum_{q\in\{1:k+1\}} \mathsf{V}_q \widehat{\mathcal{R}}(\psi_q)\circ T_n^{-1}$ for all $v := \sum_{q\in\{1:k+1\}} \mathsf{V}_q\psi_q \circ T_n^{-1}$ and all $n \in \mathcal{N}_\tau$, where $\{\psi_q\}_{q\in\{1:k+1\}}$ is a basis for $\mathbb{P}_k(\widehat{J};\mathbb{R})$. Accept as a fact that $\|v\|_{L^\infty(J_n;V')} \le 2^{2-\frac{1}{p}}\|\partial_t\mathcal{R}_n(v)\|_{L^p(J_n;V')}$ for all $p \in [1,\infty]$ and all $v \in \mathbb{P}_k(J_n;V')$ (see Holm and Wihler [27, Prop. 1]). Prove that $\|v\|_{L^2(J_n;L)} \le (2\tau)^{\frac{1}{2}}\|\partial_t\mathcal{R}_n(v)\|^{\frac{1}{2}}_{L^2(J_n;V')}\|v\|^{\frac{1}{2}}_{L^2(J_n;V)}$ for all $v \in \mathbb{P}_k(J_n;V)$ and all $n \in \mathcal{N}_\tau$. (*Hint*: $\|\phi\|_L^2 \le \|\phi\|_{V'}\|\phi\|_V$ for all $\phi \in V$.)

Exercise 69.5 (dG(1)). Assume that a is time-independent. (i) Verify that the dG(1) scheme amounts to

$$\begin{pmatrix} \frac{9}{8}\mathcal{M} & \frac{3}{8}\mathcal{M} \\ -\frac{9}{8}\mathcal{M} & \frac{5}{8}\mathcal{M}\end{pmatrix}\begin{pmatrix}\mathsf{U}^{n,1}\\ \mathsf{U}^{n,2}\end{pmatrix} + \tau\begin{pmatrix}\frac{3}{4}\mathcal{A}\mathsf{U}^{n,1}\\ \frac{1}{4}\mathcal{A}\mathsf{U}^{n,2}\end{pmatrix} = \begin{pmatrix}\frac{3}{2}\mathcal{M}\mathsf{U}^{n-1}\\ -\frac{1}{2}\mathcal{M}\mathsf{U}^{n-1}\end{pmatrix} + \tau\begin{pmatrix}\frac{3}{4}\mathsf{F}^{n,1}\\ \frac{1}{4}\mathsf{F}^{n,2}\end{pmatrix},$$

and $\mathsf{U}^n = \mathsf{U}^{n,2}$, where $\mathsf{U}^{n,1}$ and $\mathsf{U}^{n,2}$ are the coordinate vectors of the discrete solution at $t_{n-1}+\frac{1}{3}\tau$ and at t_n, respectively. (*Hint*: use the Lagrange interpolation polynomials associated with the two Gauss–Radau nodes $\xi_1 := -\frac{1}{3}$ and $\xi_2 := 1$.) (ii) Using the same notation as above, write the scheme in IRK form. (*Hint*: see (69.22) and (69.24).)

Exercise 69.6 (IRK final stage). The objective of this exercise is to prove the assertions made in Remark 69.13. (i) Show that for every s-stage IRK scheme, the update u_h^n is given by $u_h^n = \alpha_0 u_h^{n-1} + \sum_{p\in\{1:s\}} \alpha_p u_h^{n,p}$, where $\alpha_p := \sum_{q\in\{1:s\}} b_q(a^{-1})_{qp}$, $\alpha_0 := 1 - \sum_{p\in\{1:s\}} \alpha_p$, and $(a^{-1})_{pq}$ are the coefficients of the inverse of the Butcher matrix $(a_{pq})_{p,q\in\{1:s\}}$. (ii) Show that for the Radau IIA IRK scheme, $\alpha_p = 0$ for all $p \in \{0:s-1\}$ and $\alpha_s = 1$.

Exercise 69.7 (Π_τ^k). (i) Prove the uniform stability of Π_n^k in $L^\infty(J_n;Z)$ with $Z \subseteq L$. (*Hint*: map to the reference interval $\widehat{J}$.) Prove (69.27). (*Hint*: accept as a fact that the standard polynomial approximation properties in Sobolev spaces extend to Bochner spaces.) (ii) Build the operator Π_n^k with $Z := V'$ as in Remark 69.17. (*Hint*: use the Riesz–Fréchet operator $J^{\mathrm{RF}} : L^2(J_n;V) \to (L^2(J_n;V))' = L^2(J_n;V')$.) Adapt the identity in Lemma 69.16 to the case $Z := V'$. (*Hint*: invoke the integration by parts formula (64.7).) Prove a stability estimate for Π_n^k in $L^\infty(J_n;V')$. (iii) Let $\Pi_h \in \mathcal{L}(V;V_h)$. Show that $\delta := \Pi_\tau^k(\Pi_h(v)) - \Pi_h(\Pi_\tau^k(v)) = 0$ for all $v \in H^1(J;V)$. (*Hint*: show that $\delta(t_n) = 0$ for all $n \in \overline{\mathcal{N}}_\tau$ and that $\int_{J_n}(\delta,q)_L \mathrm{d}t = 0$ for all $q \in \mathbb{P}_{k-1}(J_n;V_h)$ and all $n \in \mathcal{N}_\tau$.)

Exercise 69.8 (Symmetrization). Let $\widehat{\mathcal{R}}$ be defined in Exercise 69.4(iii). (i) Prove that $\mathbb{B}_{pq} = \int_{-1}^1 \widehat{\mathcal{R}}(\psi_q)'\psi_p \,\mathrm{d}s$, $(\mathbb{B}+\mathbb{B}^\mathsf{T})_{pq} = \psi_q(-1)\psi_p(-1) + \psi_q(1)\psi_p(1)$, $(\mathbb{B}^\mathsf{T}\mathbb{M}^{-1}\mathbb{B})_{pq} = \int_{-1}^1 \widehat{\mathcal{R}}(\psi_q)'\widehat{\mathcal{R}}(\psi_p)' \,\mathrm{d}s$ for all $p,q \in \{1:m\}$. (*Hint*: use Exercise 28.1.) (ii) Set $\hat{\boldsymbol{\mathcal{S}}}_\flat := \frac{1}{\tau}(\mathcal{M}\mathcal{A}^{-1}\mathcal{M})\otimes(\mathbb{B}^\mathsf{T}\mathbb{M}^{-1}\mathbb{B}) + \tau\mathcal{A}\otimes\mathbb{M}$. Prove that $\mathsf{V}^\mathsf{T}\hat{\boldsymbol{\mathcal{S}}}_\flat\mathsf{V} \le \mathsf{V}^\mathsf{T}\hat{\boldsymbol{\mathcal{S}}}\mathsf{V} \le 2\mathsf{V}^\mathsf{T}\hat{\boldsymbol{\mathcal{S}}}_\flat\mathsf{V}$ for all $\mathsf{V} \in \mathbb{R}^{Im}$. (*Hint*: note that $\mathsf{V}^\mathsf{T}(\mathcal{M}\otimes\mathbb{B})\mathsf{V} = \mathsf{Y}^\mathsf{T}(\mathcal{A}^{-1}\otimes\mathbb{M}^{-1})\mathsf{Z}$ with $\mathsf{Y} := (\mathcal{A}\otimes\mathbb{M})\mathsf{V}$ and $\mathsf{Z} := (\mathcal{M}\otimes\mathbb{B})\mathsf{V}$ and apply the Cauchy–Schwarz and Young's inequalities.) (iii) Verify that $\hat{\boldsymbol{\mathcal{S}}}$ is the stiffness matrix associated with the minimization of the residual norm $\|A_h^{-1}(\partial_t\mathcal{R}_n(v_{h\tau}))+v_{h\tau}\|^2_{L^2(J_n;V_h)}$. (*Hint*: use again Exercise 28.1.) (iv) Compute the matrix $\hat{\boldsymbol{\mathcal{S}}}$ for $k := 1$. (*Hint*: see Exercise 69.5.)

Solution to exercises

Exercise 69.1 (Integral identities). (69.11a) follows from the fact that the discrete measure $\mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t)$ samples at the interpolation nodes of $\mathcal{I}_k^{\mathrm{GR}}$, and (69.11b) follows from (69.11a) once we observe that $\int_J (p, \mathcal{I}_k^{\mathrm{GR}}(w))_L \mathrm{d}t = \int_J (p, \mathcal{I}_k^{\mathrm{GR}}(w))_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t)$ because the quadrature is of order $2k$.

Exercise 69.2 (Equivalence with Radau IIA IRK). Let $u_{h\tau} \in X_{h\tau} := P_k^{\mathrm{b}}(\overline{J}_\tau;V_h)$ and assume that $\{u_h^{n,i} := u_{h\tau}(t_{n,i})\}_{i\in\{1:k+1\}}$ solves (69.22) with $s := k+1$ for all $n \in \mathcal{N}_\tau$. Let us define $v_{h\tau} \in P_{k+1}^{\mathrm{g}}(\overline{J}_\tau;V_h)$ by setting $v_h(0) := u_{h\tau}(0)$ and for all $t \in J_n$,

$$v_{h\tau}(t) = u_h^{n-1} + \tau \sum_{j\in\{1:k+1\}} \frac{1}{2}\int_{-1}^{T_n^{-1}(t)} \mathcal{L}_j(\xi)\,\mathrm{d}\xi\big(f_h(t_{n,j}) - \mathcal{A}_h(t_{n,j})(u_h^{n,j})\big).$$

Observe that $v_{h\tau} \in P_{k+1}^{\mathrm{b}}(\overline{J}_\tau;V_h)$ since $\mathcal{L}_i \in \mathbb{P}_k(\widehat{J};\mathbb{R})$. Moreover, $v_{h\tau}(t_{n-1}^+) = u_h^{n-1} = \mathcal{R}_\tau(u_{h\tau})(t_{n-1})$ and (69.22) together with (69.23) implies that $v_{h\tau}(t_{n,i}) = u_h^{n,i} = \mathcal{R}_\tau(u_{h\tau})(t_{n,i})$ for all $i \in \{1:k+1\}$ since $T_n^{-1}(t_{n,j}) = \xi_j$ for all $j \in \{1:k+1\}$. This proves that $v_{h\tau} = \mathcal{R}_\tau(u_{h\tau}) \in P_{k+1}^{\mathrm{g}}(\overline{J}_\tau;V_h)$. We

then infer that

$$\begin{aligned}\partial_t\mathcal{R}_\tau(u_{h\tau})(t_{n,i}) &= \tau\sum_{j\in\{1:k+1\}}\frac{1}{2}\frac{2}{\tau}\mathcal{L}_j(T_n^{-1}(t_{n,i}))\big(f_h(t_{n,j})-\mathcal{A}_h(t_{n,j})(u_h^{n,j})\big)\\ &= f_h(t_{n,i})-\mathcal{A}_h(t_{n,i})(u_h^{n,i}).\end{aligned}$$

This shows that $u_{h\tau}$ solves (69.20). Proposition 69.7 shows that $u_{h\tau}$ solves (69.16).

Exercise 69.3 (Poincaré in time). The assumption $v\in H^1(J_n;H)$ implies that v has a continuous representative in $C^0(\overline{J}_n;H)$ (see Lemma 64.37 and Remark 64.38). Using the hint, we have for all $t\in J_n$,

$$v(t)=v(t_{n-1}^+)+\int_{t_{n-1}}^t\partial_t v\mathrm{d}t.$$

The triangle inequality followed by the Cauchy–Schwarz inequality yields

$$\begin{aligned}\|v(t)\|_H &\le \|v(t_{n-1}^+)\|_H+\int_{t_{n-1}}^t\|\partial_t v\|_H\mathrm{d}t\\ &\le \|v(t_{n-1}^+)\|_H+(t-t_{n-1})^{\frac12}\bigg(\int_{t_{n-1}}^t\|\partial_t v\|_H^2\mathrm{d}t\bigg)^{\frac12}.\end{aligned}$$

Young's inequality and the fact that $(t_{n-1},t)\subset J_n$ imply that

$$\|v(t)\|_H^2\le 2\|v(t_{n-1}^+)\|_H^2+2(t-t_{n-1})\|\partial_t v\|_{L^2(J_n;H)}^2.$$

The result follows by integrating this inequality over J_n.

Exercise 69.4 (Time reconstruction). (i) Assume that R_τ is defined as in Remark 69.9 with $\theta_{k+1}(s):=\prod_{l\in\{1:k+1\}}\frac{\xi_l-s}{\xi_l+1}$. Then, for all $n\in\mathcal{N}_\tau$, we have $R_\tau(v_{h\tau})(t_{n-1}^+)=v_{h\tau}(t_{n-1}^+)-v_{h\tau}(t_{n-1}^+)+v_{h\tau}(t_{n-1})=v_{h\tau}(t_{n-1})$. Moreover, for all $l\in\{1:k+1\}$, we have $R_\tau(v_{h\tau})(t_{n,l})=v_{h\tau}(t_{n,l})$. Hence, we obtain the same operator as in Definition 69.5.
(ii) Let us set $\xi_{k+1}:=\frac{(-1)^k}{2}(L_k-L_{k+1})$ and $\delta:=\xi_{k+1}-\theta_{k+1}$. By definition, δ vanishes at $s=-1$. Moreover, since the Gauss–Radau quadrature using $(k+1)$ points is of order $2k$ (see Proposition 6.7), we infer that for all $q\in\mathbb{P}_k(\widehat{J};\mathbb{R})$,

$$\int_{\widehat{J}}\theta_{k+1}(s)q'(s)\,\mathrm{d}s=\sum_{l\in\{1:k+1\}}\omega_l\theta_{k+1}(\xi_l)q'(\xi_l)=0.$$

In addition, we have $\int_{\widehat{J}}\xi_{k+1}(s)q'(s)\,\mathrm{d}s=0$ owing to the L^2-orthogonality of the Legendre polynomials which implies that $\int_{\widehat{J}}L_k(s)q'(s)\,\mathrm{d}s=0=\int_{\widehat{J}}L_{k+1}(s)q'(s)\,\mathrm{d}s$. Using integration by parts and the fact that $\delta(1)=\xi_{k+1}(1)-\xi_{k+1}(1)=0-0=0$ (recall that $L_k(-1)=(-1)^k$), we infer that

$$\int_{\widehat{J}}\delta'(s)q(s)\,\mathrm{d}s=-\int_{\widehat{J}}\delta(s)q'(s)\,\mathrm{d}s+\big[\delta(s)q(s)\big]_{-1}^1=0.$$

Since δ' is in $\mathbb{P}_k(\widehat{J};\mathbb{R})$ and q is arbitrary in $\mathbb{P}_k(\widehat{J};\mathbb{R})$, the above identity shows that δ is constant, and since $\delta(-1)=0$ as shown above, we conclude that δ vanishes identically.
(iii) Let $v\in L^2(J_n;V)$. Recall that $V\hookrightarrow L\equiv L'\hookrightarrow V'$. Combining the hint with the Cauchy–Schwarz inequality in time leads to

$$\|v\|_{L^2(J_n;L)}\le\|v\|_{L^2(J_n;V')}^{\frac12}\|v\|_{L^2(J_n;V)}^{\frac12}.$$

This implies that

$$\|v\|_{L^2(J_n;L)} \le \tau^{\frac14}\|v\|^{\frac12}_{L^\infty(J_n;V')}\|v\|^{\frac12}_{L^2(J_n;V)}.$$

Applying the inverse estimate from [27] with $p := 2$ leads to

$$\|v\|_{L^\infty(J_n;V')} \le 2\tau^{\frac12}\|\partial_t\mathcal{R}_n(v)\|_{L^2(J_n;V')}.$$

Combining these bounds proves the claim.

Exercise 69.5 (dG(1)). (i) The Gauss–Radau nodes are $\xi_1 := -\frac13$ and $\xi_2 := 1$ and the corresponding weights are $\omega_1 := \frac32$, $\omega_2 := \frac12$ (see Table 6.1). The Lagrange interpolation polynomials are

$$\mathcal{L}_1(s) = \frac34(1-s), \qquad \mathcal{L}_2(s) = \frac34\Big(s+\frac13\Big).$$

Using these two polynomials, we have

$$(\psi_1(-1),\psi_2(-1)) = (\tfrac32,-\tfrac12), \qquad (\psi_1(1),\psi_2(1)) = (0,1),$$

and the matrices $\mathbb{B},\mathbb{M}\in\mathbb{R}^{2\times2}$ become

$$\mathbb{B} = \begin{pmatrix} \frac98 & \frac38 \\ -\frac98 & \frac58 \end{pmatrix}, \qquad \mathbb{M} = \begin{pmatrix} \frac34 & 0 \\ 0 & \frac14 \end{pmatrix}.$$

This leads to the assertion on the dG(1) time-stepping scheme:

$$\begin{pmatrix} \frac98\mathcal{M} & \frac38\mathcal{M} \\ -\frac98\mathcal{M} & \frac58\mathcal{M} \end{pmatrix}\begin{pmatrix} \mathsf{U}^{n,1} \\ \mathsf{U}^{n,2} \end{pmatrix} + \tau\begin{pmatrix} \frac34\mathcal{A}\mathsf{U}^{n,1} \\ \frac14\mathcal{A}\mathsf{U}^{n,2} \end{pmatrix} = \begin{pmatrix} \frac32\mathcal{M}\mathsf{U}^{n-1} \\ -\frac12\mathcal{M}\mathsf{U}^{n-1} \end{pmatrix} + \tau\begin{pmatrix} \frac34\mathsf{F}^{n,1} \\ \frac14\mathsf{F}^{n,2} \end{pmatrix},$$

and we set $\mathsf{U}^n := \mathsf{U}^{n,2}$.
(ii) Using (69.24), we now write the scheme in IRK form as follows:

$$\begin{pmatrix} \mathcal{M}\mathsf{U}^{n,1} \\ \mathcal{M}\mathsf{U}^{n,2} \end{pmatrix} + \tau\begin{pmatrix} \frac{5}{12}\mathcal{A} & -\frac{1}{12}\mathcal{A} \\ \frac34\mathcal{A} & \frac14\mathcal{A} \end{pmatrix}\begin{pmatrix} \mathsf{U}^{n,1} \\ \mathsf{U}^{n,2} \end{pmatrix} = \begin{pmatrix} \mathcal{M}\mathsf{U}^{n-1} \\ \mathcal{M}\mathsf{U}^{n-1} \end{pmatrix} + \tau\begin{pmatrix} \frac{5}{12}\mathsf{F}^{n,1} - \frac{1}{12}\mathsf{F}^{n,2} \\ \frac34\mathsf{F}^{n,1} + \frac14\mathsf{F}^{n,2} \end{pmatrix},$$

and we set $\mathsf{U}^n := \mathsf{U}^{n,2}$. Notice that the two linear systems are indeed equivalent.

Exercise 69.6 (IRK final stage). (i) Recalling (69.22), let us set $y_i := \frac{1}{\tau}(u_h^{n,i} - u_h^{n-1})$, $z_j := f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})$, so that we have

$$y_i = \sum_{j\in\{1:s\}} a_{ij}z_j, \qquad \forall i\in\{1:s\}.$$

This implies that

$$z_i = \sum_{j\in\{1:s\}} (a^{-1})_{ij}y_j, \qquad \forall i\in\{1:s\}.$$

Recalling that $u_h^n := u_h^{n-1} + \tau\sum_{j\in\{1:s\}} b_j\big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})\big)$, we infer that

$$\begin{aligned} u_h^n &:= u_h^{n-1} + \tau\sum_{j\in\{1:s\}} b_jz_j = u_h^{n-1} + \sum_{j\in\{1:s\}} b_j\sum_{i\in\{1:s\}}(a^{-1})_{ji}\tau y_i \\ &= u_h^{n-1}\Big(1 - \sum_{i\in\{1:s\}}\sum_{j\in\{1:s\}} b_j(a^{-1})_{ji}\Big) + \sum_{i\in\{1:s\}} u_h^{n,i}\sum_{j\in\{1:s\}} b_j(a^{-1})_{ji}. \end{aligned}$$

This shows that

$$u_h^n = \alpha_0 u_h^{n-1} + \sum_{p\in\{1:s\}} \alpha_p u_h^{n,p},$$

where $\alpha_p := \sum_{q\in\{1:s\}} b_q (a^{-1})_{qp}$, $\alpha_0 := 1 - \sum_{p\in\{1:s\}} \alpha_p$, and $(a^{-1})_{pq}$ are the coefficients of the inverse of the Butcher matrix $(a_{pq})_{p,q\in\{1:s\}}$.
(ii) In the case of the Radau IIA IRK scheme, we have $b_p = a_{sp}$ for all $p \in \{1:s\}$. We infer that

$$\alpha_p = \sum_{q\in\{1:s\}} b_q (a^{-1})_{qp} = \sum_{q\in\{1:s\}} a_{sq}(a^{-1})_{qp} = \delta_{sp},$$

for all $p \in \{1:s\}$. As a result, we have $\alpha_p = 0$ for all $p \in \{1:s-1\}$ and $\alpha_s = 1$. Finally, this implies that $\alpha_0 = 1 - 1 = 0$.

Exercise 69.7 (Π_τ^k). (i) To prove the uniform stability of Π_n^k in $L^\infty(J_n;Z)$, we define $\widehat{\Pi}^k : H^1(\widehat{J};Z) \to \mathbb{P}_k(\widehat{J};Z)$ s.t. for all $\widehat{v} \in H^1(\widehat{J};Z)$,

$$\widehat{\Pi}^k(v)(1) = \widehat{v}(1),$$
$$\int_{\widehat{J}} (\widehat{\Pi}^k(\widehat{v}) - \widehat{v}, \widehat{q})_L \mathrm{d}t = 0, \qquad \forall \widehat{q} \in \mathbb{P}_{k-1}(\widehat{J};Z).$$

This gives

$$\begin{aligned}\|\widehat{\Pi}^k(\widehat{v})\|_{L^\infty(\widehat{J};Z)} &\le \widehat{c}\Big(\|\widehat{v}(1)\|_Z + \sup_{\widehat{q}\in\mathbb{P}_{k-1}(\widehat{J};Z)} \frac{|\int_{J_n}(\widehat{\Pi}^k(\widehat{v}),\widehat{q})_L \mathrm{d}t|}{\|\widehat{q}\|_{L^2(\widehat{J};Z)}}\Big)\\ &\le \widehat{c}'\,\|\widehat{v}\|_{L^\infty(\widehat{J};Z)},\end{aligned}$$

where $\widehat{c}, \widehat{c}'$ are generic constants that can depend on k. Since we have $\Pi_n^k(v) = \widehat{\Pi}^k(v \circ T_n^{-1})$, the uniform stability of Π_n^k in $L^\infty(J_n;Z)$ follows readily.

We now prove the approximation property (69.27). Let $v \in W^{k+1,\infty}(J;Z)$. Since Π_τ^k leaves $\mathbb{P}_k^{\mathrm{g}}(\overline{J}_\tau;Z)$ pointwise invariant and is stable in $L^\infty(J;Z)$ uniformly w.r.t. τ, we infer that

$$\begin{aligned}\|v - \Pi_\tau^k(v)\|_{L^\infty(J;Z)} &= \inf_{q_\tau\in\mathbb{P}_k^{\mathrm{g}}(\overline{J}_\tau;Z)} \|v - q + \Pi_\tau^k(q-v)\|_{L^\infty(J;Z)}\\ &\le c \inf_{q_\tau\in\mathbb{P}_k^{\mathrm{g}}(\overline{J}_\tau;Z)} \|v - q\|_{L^\infty(J;Z)}.\end{aligned}$$

The expected estimate follows from standard polynomial approximation properties extended to Bochner Sobolev spaces.
(ii) Let $v \in H^1(J_n;V')$. We want to build $\Pi_n^k(v) \in \mathbb{P}_k(J_n;V')$ s.t.

$$\Pi_n^k(v)(t_n) = v(t_n),$$
$$\int_{J_n} \langle \Pi_n^k(v) - v, q\rangle_{V',V} \mathrm{d}t = 0, \qquad \forall q \in \mathbb{P}_{k-1}(J_n;V).$$

Let $J^{\mathrm{RF}} : L^2(J_n;V) \to (L^2(J_n;V))' = L^2(J_n;V')$ be the Riesz–Fréchet map associated with the Hilbert space $L^2(J_n;V)$ (see Theorem C.24). Since $\mathbb{P}_{k-1}(J_n;V)$ is a closed subspace of $L^2(J_n;V)$, the $L^2(J_n;V)$-orthogonal projection $\Theta_{k-1}^n : L^2(J_n;V) \to \mathbb{P}_{k-1}(J_n;V)$ is well defined. Let us set

$$\tilde{\Theta}_{k-1}^n := J^{\mathrm{RF}} \circ \Theta_{k-1}^n \circ (J^{\mathrm{RF}})^{-1} : L^2(J_n;V') \to L^2(J_n;V').$$

For all $q \in \mathbb{P}_{k-1}(J_n;V)$, we have

$$\begin{aligned}\int_{J_n} \langle \tilde{\Theta}^n_{k-1}(v), q\rangle_{V',V} \mathrm{d}t &= \int_{J_n} \langle J^{\mathrm{RF}}(\Theta^n_{k-1}((J^{\mathrm{RF}})^{-1}(v))), q\rangle_{V',V} \mathrm{d}t \\ &= \int_{J_n} (\Theta^n_{k-1}((J^{\mathrm{RF}})^{-1}(v)), q)_V \mathrm{d}t \\ &= \int_{J_n} ((J^{\mathrm{RF}})^{-1}(v), q)_V \mathrm{d}t = \int_{J_n} \langle v, q\rangle_{V',V} \mathrm{d}t.\end{aligned}$$

We then observe that for all $v \in H^1(J_n;V')$, the definition

$$\Pi^k_n(v)(t) := (v(t_n) - \tilde{\Theta}^n_{k-1}(v)(t_n))L_k(t) + \tilde{\Theta}^n_{k-1}(v)(t), \qquad \forall t \in J_n,$$

satisfies all of the above requirements. Moreover, by invoking the integration by parts formula (64.7), the identity in Lemma 69.16 becomes

$$\int_{J_n} \langle \partial_t(v - \Pi^k_\tau(v)), y_{h\tau}\rangle_{V',V} \mathrm{d}t - ([\![\Pi^k_\tau(v)]\!]_{n-1}, y_{h\tau}(t^+_{n-1}))_L = 0,$$

for all $v \in H^1(J;V)$, all $y_{h\tau} \in X_{h\tau}$, and all $n \in \mathcal{N}_\tau$. Finally, let us derive a stability estimate on Π^k_n in $L^\infty(J_n;V')$. We first notice that

$$\begin{aligned}\|\tilde{\Theta}^n_{k-1}\|_{\mathcal{L}(L^2(V');L^2(V'))} &= \|\Theta^n_{k-1} \circ (J^{\mathrm{RF}})^{-1})\|_{\mathcal{L}(L^2(J;V');L^2(J;V))} \\ &\le \|(J^{\mathrm{RF}})^{-1}\|_{\mathcal{L}(L^2(J_n;V');L^2(J;V))} = 1.\end{aligned}$$

Moreover, reasoning as above shows that there exists c s.t. for all $v \in H^1(J_n;V')$,

$$\|\Pi^k_n(v)\|_{L^\infty(J_n;V')} \le c\|v\|_{L^\infty(J_n;V')}.$$

(iii) Let $v \in H^1(J;V)$. We observe that $\Pi^k_\tau(\Pi_h(v))$ and $\Pi_h(\Pi^k_\tau(v))$ are both in $P^{\mathrm{b}}_k(\overline{J}_\tau;V_h)$. Therefore, the assertion is established provided we show that

$$\begin{aligned}&\Pi^k_\tau(\Pi_h(v))(t_n) = \Pi_h(\Pi^k_\tau(v))(t_n), && \forall n \in \overline{\mathcal{N}}_\tau, \\ &\int_{J_n} (\Pi^k_\tau(\Pi_h(v)) - \Pi_h(\Pi^k_\tau(v)), q)_L \mathrm{d}t = 0, && \forall q \in \mathbb{P}_{k-1}(J_n;V_h),\ \forall n \in \mathcal{N}_\tau.\end{aligned}$$

By definition of Π^k_τ, we have for all $n \in \overline{\mathcal{N}}_\tau$,

$$\Pi^k_\tau(\Pi_h(v))(t_n) = \Pi_h(v)(t_n) = \Pi_h(v(t_n)) = \Pi_h(\Pi^k_\tau(v)(t_n)) = \Pi_h(\Pi^k_\tau(v))(t_n).$$

Moreover, for all $q \in \mathbb{P}_{k-1}(J_n;V_h)$ and all $n \in \mathcal{N}_\tau$, since the duality between V an V' is an extension of the L-inner product, we have

$$\int_{J_n} (\Pi^k_\tau(\Pi_h(v)), q)_L \mathrm{d}t = \int_{J_n} (\Pi_h(v), q)_L \mathrm{d}t = \int_{J_n} \langle v, (\Pi_h)^*(q)\rangle_{V,V'} \mathrm{d}t,$$

where $(\Pi_h)^* : V_h \to V'$. Since $(\Pi_h)^*(q) \in \mathbb{P}_{k-1}(J_n;V')$, we infer that

$$\int_{J_n} \langle v, (\Pi_h)^*(q)\rangle_{V,V'} \mathrm{d}t = \int_{J_n} \langle \Pi^k_\tau(v), (\Pi_h)^*(q)\rangle_{V,V'} \mathrm{d}t = \int_{J_n} (\Pi_h(\Pi^k_\tau(v)), q)_L \mathrm{d}t.$$

This completes the proof.

Exercise 69.8 (Symmetrization). (i) We have

$$\begin{aligned}\int_{-1}^{1} \widehat{\mathcal{R}}(\psi_q)'\psi_p \,\mathrm{d}s &= \int_{-1}^{1} (\psi_q - \psi_q(-1)\theta_{k+1})'\psi_p \,\mathrm{d}s \\ &= \int_{-1}^{1} \psi_q'\psi_p \,\mathrm{d}s - \psi_q(-1)\int_{-1}^{1} \theta_{k+1}'\psi_p \,\mathrm{d}s \\ &= \int_{-1}^{1} \psi_q'\psi_p \,\mathrm{d}s + \psi_q(-1)\psi_p(-1) = \mathbb{B}_{pq},\end{aligned}$$

where we integrated by parts and used that $\int_{-1}^{1} \theta_{k+1} r \,\mathrm{d}s = 0$ for all $r \in \mathbb{P}_{k-1}(\widehat{J};\mathbb{R})$, $\theta_{k+1}(-1) = 1$, and $\theta_{k+1}(1) = 0$ to pass from the second to the third line. Furthermore, we have

$$\begin{aligned}(\mathbb{B} + \mathbb{B}^{\mathsf{T}})_{pq} &= \int_{-1}^{1} (\psi_q'\psi_p + \psi_p'\psi_q) \,\mathrm{d}s + 2\psi_q(-1)\psi_p(-1) \\ &= \psi_q(-1)\psi_p(-1) + \psi_q(1)\psi_p(1),\end{aligned}$$

since $\psi_q'\psi_p + \psi_p'\psi_q = (\psi_q\psi_p)'$. Finally, the identity

$$(\mathbb{B}^{\mathsf{T}}\mathbb{M}^{-1}\mathbb{B})_{pq} = \int_{-1}^{1} \widehat{\mathcal{R}}(\psi_q)'\widehat{\mathcal{R}}(\psi_p)' \,\mathrm{d}s$$

has been shown in Exercise 28.1 with the operator $Z : \mathbb{P}_k(\widehat{J};\mathbb{R}) \to \mathbb{P}_k(\widehat{J};\mathbb{R})$ s.t. $Z(r) = (\widehat{\mathcal{R}}(r))'$. (ii) From Step (i), we observe that the matrix $(\mathbb{B} + \mathbb{B}^{\mathsf{T}})$ is (symmetric) positive semidefinite. Indeed, setting $\Psi(\pm 1) = (\psi_1(\pm 1), \dots, \psi_{k+1}(\pm 1))^{\mathsf{T}} \in \mathbb{R}^{k+1}$, we have $(\mathbb{B}+\mathbb{B}^{\mathsf{T}}) = \Psi(-1)\otimes\Psi(-1)^{\mathsf{T}} + \Psi(1)\otimes\Psi(1)^{\mathsf{T}}$ and $\mathsf{X}^{\mathsf{T}}(\mathbb{B} + \mathbb{B}^{\mathsf{T}})\mathsf{X} = (\mathsf{X}^{\mathsf{T}}\Psi(-1))^2 + (\mathsf{X}^{\mathsf{T}}\Psi(1))^2$. Moreover, the matrix $\mathcal{M}$ is positive definite. This implies that $\mathcal{M} \otimes (\mathbb{B} + \mathbb{B}^{\mathsf{T}})$ is symmetric positive semidefinite. After noticing that $\hat{\boldsymbol{\mathcal{S}}} - \hat{\boldsymbol{\mathcal{S}}}_\flat = \mathcal{M} \otimes (\mathbb{B}+\mathbb{B}^{\mathsf{T}})$, we then infer that $\mathsf{V}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{S}}}_\flat\mathsf{V} \le \mathsf{V}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{S}}}\mathsf{V}$. Since the matrix $\mathcal{M}$ is symmetric and since the matrix $\mathcal{A} \otimes \mathbb{M}$ is symmetric positive definite, the Cauchy–Schwarz inequality followed by Young's inequality applied to $\mathsf{Y} := (\mathcal{A} \otimes \mathbb{M})\mathsf{V}$ and $\mathsf{Z} := (\mathcal{M} \otimes \mathbb{B})\mathsf{V}$ implies that

$$\begin{aligned}\mathsf{V}^{\mathsf{T}}(\mathcal{M} \otimes (\mathbb{B} + \mathbb{B}^{\mathsf{T}}))\mathsf{V} &= 2\mathsf{V}^{\mathsf{T}}(\mathcal{M} \otimes \mathbb{B})\mathsf{V} \\ &= 2\mathsf{V}^{\mathsf{T}}(\mathcal{A} \otimes \mathbb{M})(\mathcal{A}^{-1} \otimes \mathbb{M}^{-1})(\mathcal{M} \otimes \mathbb{B})\mathsf{V} \\ &= 2\mathsf{Y}^{\mathsf{T}}(\mathcal{A}^{-1} \otimes \mathbb{M}^{-1})\mathsf{Z} \\ &\le \tau\mathsf{Y}^{\mathsf{T}}(\mathcal{A}^{-1} \otimes \mathbb{M}^{-1})\mathsf{Y} + \tau^{-1}\mathsf{Z}^{\mathsf{T}}(\mathcal{A}^{-1} \otimes \mathbb{M}^{-1})\mathsf{Z} \\ &= \tau\mathsf{V}^{\mathsf{T}}(\mathcal{A} \otimes \mathbb{M})\mathsf{V} + \tau^{-1}\mathsf{V}^{\mathsf{T}}((\mathcal{M}\mathcal{A}^{-1}\mathcal{M}) \otimes \mathbb{B}^{\mathsf{T}}\mathbb{M}^{-1}\mathbb{B}))\mathsf{V} \\ &= \mathsf{V}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{S}}}_\flat\mathsf{V}.\end{aligned}$$

Hence, we have

$$\mathsf{V}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{S}}}\mathsf{V} = \mathsf{V}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{S}}}_\flat\mathsf{V} + \mathsf{V}^{\mathsf{T}}(\mathcal{M} \otimes (\mathbb{B} + \mathbb{B}^{\mathsf{T}}))\mathsf{V} \le 2\mathsf{V}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{S}}}_\flat\mathsf{V}.$$

(iii) We have

$$\begin{aligned}\boldsymbol{\mathcal{S}}_{jq,ip} &= \int_{J_n} (\partial_t\mathcal{R}_n(\varphi_j\psi_q) + A_h(\varphi_j\psi_q), \varphi_i\psi_p)_L \mathrm{d}t \\ &= (A_h^{-1}(\partial_t\mathcal{R}_n(\varphi_j\psi_q)) + \varphi_j\psi_q, \varphi_i\psi_p)_{L^2(J_n;V_h)},\end{aligned}$$

for all $i, j \in \{1{:}I\}$ and all $p, q \in \{1{:}m\}$. Moreover, the mass matrix of the inner product $(\cdot,\cdot)_{L^2(J_n;V_h)}$ is $\mathcal{A} \otimes \tau\mathbb{M}$. We conclude by invoking the result of Exercise 28.1.
(iv) For $k := 1$, an explicit computation shows that

$$\mathbb{B}^{\mathsf{T}}\mathbb{M}^{-1}\mathbb{P} = \begin{pmatrix} \frac{27}{4} & -\frac{9}{4} \\ -\frac{9}{4} & \frac{7}{4} \end{pmatrix}, \qquad \mathbb{B} + \mathbb{B}^{\mathsf{T}} = \begin{pmatrix} \frac{9}{4} & -\frac{3}{4} \\ -\frac{3}{4} & \frac{5}{4} \end{pmatrix}.$$

In conclusion, the preconditioned symmetric dG(1) system matrix takes the following form:

$$\hat{\boldsymbol{\mathcal{S}}} = \tfrac{1}{\tau}(\mathcal{M}\mathcal{A}^{-1}\mathcal{M}) \otimes \begin{pmatrix} \frac{27}{4} & -\frac{9}{4} \\ -\frac{9}{4} & \frac{7}{4} \end{pmatrix} + \mathcal{M} \otimes \begin{pmatrix} \frac{9}{4} & -\frac{3}{4} \\ -\frac{3}{4} & \frac{5}{4} \end{pmatrix} + \tau\mathcal{A} \begin{pmatrix} \frac{3}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}.$$

Chapter 70

Continuous Petrov–Galerkin in time

Exercises

Exercise 70.1 (Interpolation operators). (i) Let $\mathcal{I}_{k-1}^{\mathrm{GL}}$ be the Lagrange interpolation operator defined in (70.2) using $Z := L$. Prove that

$$\int_J (p, \mathcal{I}_{k-1}^{\mathrm{GL}}(w))_L \mathrm{d}t = \int_J (p,w)_L \, \mu_k^{\mathrm{GL}}(\mathrm{d}t), \tag{70.1a}$$

$$\int_J (v, \mathcal{I}_{k-1}^{\mathrm{GL}}(w))_L \, \mu_k^{\mathrm{GL}}(\mathrm{d}t) = \int_J (v,w)_L \, \mu_k^{\mathrm{GL}}(\mathrm{d}t), \tag{70.1b}$$

for all $p \in P_k^{\mathrm{b}}(J_\tau; L)$ and all $v, w \in L^2(J;L)$. (ii) Let $Z \subseteq L$. Prove that the restriction of $\mathcal{I}_{k-1}^{\mathrm{GL}}$ to $P_k^{\mathrm{g}}(\overline{J}_\tau; Z)$ coincides with the $L^2(J;Z)$-orthogonal projection onto $P_{k-1}^{\mathrm{b}}(J_\tau; Z)$. (iii) Prove (70.5).

Exercise 70.2 (Equivalence with KB IRK). Prove the converse assertion in Lemma 70.5. $\big($*Hint*: show that $u_{h\tau}(t) = u_h^{n-1} + \tau \sum_{j\in\{1:k\}} \frac{1}{2} \int_{-1}^{T_n^{-1}(t)} \mathcal{L}_j(\xi) \, \mathrm{d}\xi \big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j}) \big)$ for all $t \in J_n$ and all $n \in \mathcal{N}_\tau$.$\big)$

Exercise 70.3 (Butcher simplifying assumptions). Let $s \in \mathbb{N} \setminus \{0\}$ and let $\{c_i\}_{i\in\{1:s\}}$ be s distinct points in $[0,1]$. Let $\xi_i := 2c_i - 1$ and $\mathcal{L}_i(\xi) := \prod_{j\in\{1:s\}\setminus\{i\}} \frac{\xi-\xi_j}{\xi_i-\xi_j}$ for all $i \in \{1{:}s\}$. Let $a_{ij} := \frac{1}{2}\int_{-1}^{2c_i-1} \mathcal{L}_j(\xi)\,\mathrm{d}\xi$, $b_i := \frac{1}{2}\int_{-1}^{1} \mathcal{L}_i(\xi)\,\mathrm{d}\xi$ for all $i \in \{1{:}s\}$. (i) Show that the set $\{\xi_i, 2b_i\}_{i\in\{1:s\}}$ is a quadrature of order $k_{\mathcal{Q}} \geq s-1$ over the interval $[-1,1]$ (see Definition 6.4). (*Hint*: observe that $p = \sum_{i\in\{1:s\}} p(\xi_i)\mathcal{L}_i$ for all $p \in \mathbb{P}_{s-1}(\widehat{J};\mathbb{R})$.) (ii) Show that for all $q \in \{1{:}s\}$,

$$\sum_{j\in\{1:s\}} a_{ij} c_j^{q-1} = \frac{c_i^q}{q}, \ \forall i \in \{1{:}s\}, \qquad \sum_{j\in\{1:s\}} b_j c_j^{q-1} = \frac{1}{q}.$$

(*Hint*: integrate $\big(\frac{1+\xi}{2}\big)^{q-1}$ over $(-1,\xi_i)$ for all $i \in \{1{:}s\}$ and over $(-1,1)$.) (iii) Assuming that $k_{\mathcal{Q}} \geq s$, show that for all $j \in \{1{:}s\}$,

$$\sum_{i\in\{1:s\}} b_i c_i^{q-1} a_{ij} = \frac{b_j}{q}(1 - c_j^q), \qquad \forall q \in \{1{:}k_{\mathcal{Q}}-s+1\}.$$

(*Hint*: integrate the polynomial $\left(\frac{1+\xi}{2}\right)^{q-1}\int_{-1}^{\xi}\mathcal{L}_j(\xi)\,\mathrm{d}\xi$ over $(-1,1)$.) *Note*: these formulae are called *Butcher's simplifying assumptions* in the ODE literature (see Butcher [9, Thm. 7], Hairer et al. [25, §II.6], [24, §IV.5, Thm. 5.1], see also the order conditions stated in Theorem 78.5).

Exercise 70.4 (cPG(k)). Assume that a is time-independent. (i) Use the IRK formalism and the tableaux in (70.15) to write the algebraic form of cPG(1) and cPG(2). (*Hint*: use the coefficients $\{\alpha_i\}_{i\in\{0:s\}}$.) (ii) Write again the cPG(1) and cPG(2) schemes in algebraic form using the formalism described in §70.3.2 and the bases from Remark 70.16. (*Hint*: for $k:=1$, it is of the form $(2\mathcal{M}+\tau\mathcal{A})\mathsf{U}^{n,1}=2\mathcal{M}\mathsf{U}^{n-1}+\tau\mathsf{F}^{n,1}$ and $\mathsf{U}^n=2\mathsf{U}^{n,1}-\mathsf{U}^{n-1}$, whereas for $k:=2$, it is of the form

$$\begin{pmatrix}\frac{3}{2} & \frac{2\sqrt{3}-3}{2}\\ -\frac{2\sqrt{3}+3}{2} & \frac{3}{2}\end{pmatrix}\begin{pmatrix}\mathcal{M}\mathsf{U}^{n,1}\\ \mathcal{M}\mathsf{U}^{n,2}\end{pmatrix}+\frac{\tau}{2}\begin{pmatrix}\mathcal{A}\mathsf{U}^{n,1}\\ \mathcal{A}\mathsf{U}^{n,2}\end{pmatrix}=\begin{pmatrix}\sqrt{3}\mathcal{M}\mathsf{U}^{n-1}+\frac{\tau}{2}\mathsf{F}^{n,1}\\ -\sqrt{3}\mathcal{M}\mathsf{U}^{n-1}+\frac{\tau}{2}\mathsf{F}^{n,2}\end{pmatrix},$$

and $\mathsf{U}^n:=\mathsf{U}^{n-1}-\sqrt{3}(\mathsf{U}^{n,1}-\mathsf{U}^{n,2})$.)

Exercise 70.5 (Π_τ^k and Π_h commute). Let $\Pi_h\in\mathcal{L}(V;V_h)$. Show that $\Pi_\tau^k(\Pi_h(v))=\Pi_h(\Pi_\tau^k(v))$ for all $v\in H^1(J;V)$. (*Hint*: use Remark 70.10 and prove that Π_h commutes with Ξ_{k-1}^{b} by introducing $(\Pi_h)^*\in\mathcal{L}(V_h;V')$.)

Solution to exercises

Exercise 70.1 (Interpolation operators). (i) The identity (70.1b) follows from the fact that the discrete measure $\mu_k^{\mathrm{GL}}(\mathrm{d}t)$ samples at the interpolation nodes of $\mathcal{I}_{k-1}^{\mathrm{GL}}$, and the identity (70.1a) follows from (70.1b) once we observe that $\int_J(p,\mathcal{I}_{k-1}^{\mathrm{GL}}(w))_L\mathrm{d}t=\int_J(p,\mathcal{I}_{k-1}^{\mathrm{GL}}(w))_L\,\mu_k^{\mathrm{GL}}(\mathrm{d}t)$ because the quadrature is of order $(2k-1)$.
(ii) Since the quadrature is of order $(2k-1)$, we have for all $v_\tau\in P_k^{\mathrm{g}}(\overline{J}_\tau;Z)$ and all $y_\tau\in P_{k-1}^{\mathrm{b}}(J_\tau;Z)$,

$$\int_J(\mathcal{I}_{k-1}^{\mathrm{GL}}(v_\tau),y_\tau)_L\mathrm{d}t=\int_J(\mathcal{I}_{k-1}^{\mathrm{GL}}(v_\tau),y_\tau)_L\,\mu_k^{\mathrm{GL}}(\mathrm{d}t)=\int_J(v_\tau,y_\tau)_L\mathrm{d}t.$$

This proves the assertion.
(iii) Let us prove (70.5). Let $v\in H^1(J;L)$ and $y_\tau\in P_{k-1}^{\mathrm{b}}(J_\tau;L)$. We observe that by construction $\mathcal{I}_k^{\mathrm{GL}+}(v)$ coincides with v at the Gauss–Legendre nodes $\{t_{n,l}\}_{l\in\{1:k\}}$ over each time interval J_n for all $n\in\mathcal{N}_\tau$. Hence, we have

$$\int_J(v,y_\tau)_L\mu_k^{\mathrm{GL}}(\mathrm{d}t)=\int_J(\mathcal{I}_k^{\mathrm{GL}+}(v),y_\tau)_L\mu_k^{\mathrm{GL}}(\mathrm{d}t).$$

But since $(\mathcal{I}_k^{\mathrm{GL}+}(v),y_\tau)_L\in\mathbb{P}_{2k-1}(J_\tau;\mathbb{R})$, the quadrature is exact, and we have

$$\int_J(v,y_\tau)_L\mu_k^{\mathrm{GL}}(\mathrm{d}t)=\int_J(\mathcal{I}_k^{\mathrm{GL}+}(v),y_\tau)_L\mathrm{d}t.$$

Exercise 70.2 (Equivalence with KB IRK). Assume that $\{u_h^{n,l}\}_{l\in\{1:s\}}$ solves (70.12) with $s:=k$ and u_h^n is given by (70.13) for all $n\in\mathcal{N}_\tau$. Let $u_{h\tau}\in P_k^{\mathrm{g}}(\overline{J}_\tau;V_h)$ be s.t. $u_{h\tau}(t_n):=u_h^n$ for all $n\in\overline{\mathcal{N}}_\tau$, and $\{u_{h\tau}(t_{n,l}):=u_h^{n,l}\}_{l\in\{1:k\}}$ for all $n\in\mathcal{N}_\tau$. Let us define $v_{h\tau}\in P_k^{\mathrm{b}}(\overline{J}_\tau;V_h)$ by setting for all $t\in J_n$ and all $n\in\mathcal{N}_\tau$,

$$v_{h\tau}(t):=u_h^{n-1}+\tau\sum_{j\in\{1:k\}}\frac{1}{2}\int_{-1}^{T_n^{-1}(t)}\mathcal{L}_j(\xi)\,\mathrm{d}\xi\big(f_h(t_{n,j})-A_h(t_{n,j})(u_h^{n,j})\big).$$

Observe that indeed $v_{h\tau} \in P_k^{\mathrm{b}}(\overline{J}_\tau; V_h)$ since $\mathcal{L}_j \in \mathbb{P}_{k-1}(\widehat{J};\mathbb{R})$ for all $j \in \{1{:}k\}$. Moreover, (70.12) together with (70.11) implies that $v_{h\tau}(t_{n,i}) = u_h^{n,i}$ for all $i \in \{1{:}k\}$. Similarly, (70.13) together with (70.11) implies that $v_{h\tau}(t_n) = u_h^n$. Finally, $v_{h\tau}(t_{n-1}^+) = u_h^{n-1}$. These arguments prove that $v_{h\tau} = u_{h\tau} \in P_k^{\mathrm{g}}(\overline{J}_\tau; V_h)$. Recalling that $T_n^{-1}(t) = 2\frac{t-t_{n-1}}{\tau} - 1$, we obtain for all $i \in \{1{:}k\}$,

$$\begin{aligned}\partial_t u_{h\tau}(t_{n,i}) &= \tau \sum_{j\in\{1:k\}} \frac{1}{2}\frac{2}{\tau}\mathcal{L}_j(T_n^{-1}(t_{n,i}))\big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})\big)\\ &= f_h(t_{n,i}) - A_h(t_{n,i})(u_h^{n,i}).\end{aligned}$$

This shows that $u_{h\tau}$ solves the cPG(k) scheme (70.9).

Exercise 70.3 (Butcher simplifying assumptions). (i) Recall that the polynomials $\{\mathcal{L}_i\}_{i\in\{1:s\}}$ are the Lagrange polynomials associated with the nodes $\{\xi_i\}_{i\in\{1:s\}}$. Hence they form a basis of $\mathbb{P}_{s-1}(\widehat{J};\mathbb{R})$. This implies that for all $p = \sum_{i\in\{1:s\}} p(\xi_i)\mathcal{L}_i \in \mathbb{P}_{s-1}(\widehat{J};\mathbb{R})$, we have

$$\int_{-1}^1 p(\xi)\,\mathrm{d}\xi = \sum_{i\in\{1:s\}} p(\xi_i)\int_{-1}^1 \mathcal{L}_i(\xi)\,\mathrm{d}\xi = \sum_{i\in\{1:s\}} 2b_i p(\xi_i).$$

This shows that the set $\{\xi_i, 2b_i\}_{i\in\{1:s\}}$ is a quadrature of order $k_{\mathcal{Q}} \ge s-1$ (see Definition 6.4).
(ii) Since the polynomials $\{\mathcal{L}_i\}_{i\in\{1:s\}}$ are the Lagrange polynomials associated with the nodes $\{\xi_i\}_{i\in\{1:s\}}$ and they form a basis of $\mathbb{P}_{s-1}(\widehat{J};\mathbb{R})$, we have for all $q \in \{1{:}s-1\}$,

$$\left(\frac{1+\xi}{2}\right)^{q-1} = \sum_{j\in\{1:s\}} \left(\frac{1+\xi_j}{2}\right)^{q-1}\mathcal{L}_j(\xi).$$

Integrating over $(-1,\xi_i)$ for all $i \in \{1{:}s\}$ and using the definition of c_i gives

$$\begin{aligned}\frac{2c_i^q}{q} = \frac{2}{q}\left(\frac{1+\xi_i}{2}\right)^q &= \int_{-1}^{\xi_i}\left(\frac{1+\xi}{2}\right)^{q-1}\mathrm{d}\xi\\ &= \sum_{j\in\{1:s\}}\left(\frac{1+\xi_j}{2}\right)^{q-1}\int_{-1}^{\xi_i}\mathcal{L}_j(\xi)\,\mathrm{d}\xi.\end{aligned}$$

Recalling that $c_j = \frac{1+\xi_j}{2}$ and $a_{ij} = \frac{1}{2}\int_{-1}^{\xi_i}\mathcal{L}_j(\xi)\,\mathrm{d}\xi$, we obtain

$$\frac{c_i^q}{q} = \sum_{j\in\{1:s\}} a_{ij}c_j^{q-1}, \qquad \forall i \in \{1{:}s\}.$$

Now, we integrate over $(-1,1)$ and get

$$\frac{2}{q} = \int_{-1}^1\left(\frac{1+\xi}{2}\right)^{q-1}\mathrm{d}\xi = \sum_{j\in\{1:s\}}\left(\frac{1+\xi_j}{2}\right)^{q-1}\int_{-1}^1\mathcal{L}_j(\xi)\,\mathrm{d}\xi.$$

Recalling that $b_j = \frac{1}{2}\int_{-1}^1\mathcal{L}_j(\xi)\,\mathrm{d}\xi$ for all $j \in \{1{:}s\}$, we obtain

$$\frac{1}{q} = \sum_{j\in\{1:s\}} b_j c_j^{q-1}.$$

(iii) Let $k_{\mathcal{Q}} \geq s$ be the order of the quadrature. Let $q \in \{1\!:\!k_{\mathcal{Q}}-s+1\}$. Since $\left(\frac{1+\xi}{2}\right)^{q-1} \int_{-1}^{\xi} \mathcal{L}_j(\xi)\,\mathrm{d}\xi$ is a polynomial of degree $q+s-1 \leq k_{\mathcal{Q}}$, we infer that

$$\begin{aligned}\frac{1}{4}\int_{-1}^{1}\left(\frac{1+\xi}{2}\right)^{q-1}\int_{-1}^{\xi}\mathcal{L}_j(\zeta)\,\mathrm{d}\zeta\,\mathrm{d}\xi &= \frac{1}{4}\sum_{i\in\{1:s\}} 2b_i\left(\frac{1+\xi_i}{2}\right)^{q-1}\int_{-1}^{\xi_i}\mathcal{L}_j(\zeta)\,\mathrm{d}\zeta\\ &= \sum_{i\in\{1:s\}} b_i c_i^{q-1} a_{ij}.\end{aligned}$$

Moreover, integrating by parts, we obtain

$$\begin{aligned}\frac{1}{4}\int_{-1}^{1}\left(\frac{1+\xi}{2}\right)^{q-1}\int_{-1}^{\xi}\mathcal{L}_j(\zeta)\,\mathrm{d}\zeta\,\mathrm{d}\xi = &-\frac{1}{4}\int_{-1}^{1}\frac{2}{q}\left(\frac{1+\xi}{2}\right)^{q}\mathcal{L}_j(\xi)\,\mathrm{d}\xi\\ &+\frac{1}{4}\left[\frac{2}{q}\left(\frac{1+\xi}{2}\right)^{q}\int_{-1}^{\xi}\mathcal{L}_j(\zeta)\,\mathrm{d}\zeta\right]_{-1}^{1}.\end{aligned}$$

Since the degree of $\left(\frac{1+\xi}{2}\right)^{q}\mathcal{L}_j(\xi)$ is $q+s-1 \leq k_{\mathcal{Q}}$, we infer that

$$\frac{1}{4}\int_{-1}^{1}\left(\frac{1+\xi}{2}\right)^{q-1}\int_{-1}^{\xi}\mathcal{L}_j(\zeta)\,\mathrm{d}\zeta\,\mathrm{d}\xi = -\frac{1}{4}\frac{2}{q}c_j^q 2b_j + \frac{1}{4}\frac{2}{q}2b_j.$$

In conclusion, we have established that for all $j \in \{1\!:\!s\}$,

$$\sum_{i\in\{1:s\}} b_i c_i^{q-1} a_{ij} = \frac{b_j}{q}(1-c_j^q).$$

Exercise 70.4 (cPG(k)). (i) Let us start with $k = s := 1$. The corresponding Butcher tableau in (70.15) gives

$$\begin{aligned}&\mathcal{M}\mathsf{U}^{n,1} + \frac{1}{2}\tau\mathcal{A}\mathsf{U}^{n,1} = \mathcal{M}\mathsf{U}^{n-1} + \frac{1}{2}\tau\mathsf{F}^{n,1},\\ &\mathcal{M}\mathsf{U}^{n} = \mathcal{M}\mathsf{U}^{n-1} + \tau\mathsf{F}^{n,1} - \tau\mathcal{A}\mathsf{U}^{n,1},\end{aligned}$$

with $\mathsf{F}_i^{n,1} := \langle f(t_{n,1}), \varphi_i\rangle_{V',V}$ for all $i \in I$ and $t_{n,1} := \frac{t_{n-1}+t_n}{2}$. We can eliminate the intermediate state $\mathsf{U}^{n,1}$. Indeed, according to (70.14) and the coefficients $\{\alpha_i\}_{i\in\{0:s\}}$ given below (70.15) for $s := 1$, we have $\mathsf{U}^n = -\mathsf{U}^{n-1} + 2\mathsf{U}^{n,1}$. Hence, $\mathsf{U}^{n,1} = \frac{1}{2}(\mathsf{U}^n + \mathsf{U}^{n-1})$. After inserting this expression into the first equation, we obtain the Crank–Nicolson time-stepping scheme:

$$\mathcal{M}(\mathsf{U}^n - \mathsf{U}^{n-1}) + \frac{1}{2}\tau\mathcal{A}(\mathsf{U}^n + \mathsf{U}^{n-1}) = \tau\mathsf{F}^{n,1}.$$

Now, for $k = s := 2$, we have (using (70.14) and the coefficients $\{\alpha_i\}_{i\in\{0:s\}}$ given below (70.15))

$$\begin{aligned}&\mathcal{M}\mathsf{U}^{n,1} + \frac{1}{4}\tau\mathcal{A}\mathsf{U}^{n,1} + \frac{3-2\sqrt{3}}{12}\tau\mathcal{A}\mathsf{U}^{n,2} = \mathcal{M}\mathsf{U}^{n-1} + \frac{1}{4}\tau\mathsf{F}^{n,1} + \frac{3-2\sqrt{3}}{12}\tau\mathsf{F}^{n,2},\\ &\frac{3+2\sqrt{3}}{12}\tau\mathcal{A}\mathsf{U}^{n,1} + \mathcal{M}\mathsf{U}^{n,2} + \frac{1}{4}\tau\mathcal{A}\mathsf{U}^{n,2} = \mathcal{M}\mathsf{U}^{n-1} + \frac{3+2\sqrt{3}}{12}\tau\mathsf{F}^{n,1} + \tau\frac{1}{4}\mathsf{F}^{n,2},\\ &\mathsf{U}^n = \mathsf{U}^{n-1} - \sqrt{3}(\mathsf{U}^{n,1} - \mathsf{U}^{n,2}),\end{aligned}$$

with $\mathsf{F}_i^{n,1} := \langle f(t_{n,1}), \varphi_i\rangle_{V',V}$, $\mathsf{F}_i^{n,2} := \langle f(t_{n,2}), \varphi_i\rangle_{V',V}$ for all $i \in \{1{:}I\}$, and $t_{n,1} := \frac{t_{n-1}+t_n}{2} - \frac{\sqrt{3}}{3}\tau$, $t_{n,2} := \frac{t_{n-1}+t_n}{2} + \frac{\sqrt{3}}{3}\tau$.

(ii) We now write again the cPG(1) scheme and the cPG(2) scheme in algebraic form using the formalism described in §70.3.2 and the bases from Remark 70.16. For $k := 1$, the basis of $\mathbb{P}_1(\widehat{J};\mathbb{R})$ is $\phi_0(s) = -s$ and $\phi_1(s) = 1+s$, whereas the basis of $\mathbb{P}_0(\widehat{J};\mathbb{R})$ is $\psi_1(s) = 1$. This leads to $b_{11} = 2$, $m_{11} = 1$, $d_1 = 2$, $\alpha_0 = -1$, and $\alpha_1 = 2$. Hence, we have

$$\begin{aligned}(2\mathcal{M} + \tau\mathcal{A})\mathsf{U}^{n,1} &= 2\mathcal{M}\mathsf{U}^{n-1} + \tau\mathsf{F}^{n,1},\\ \mathsf{U}^n &= 2\mathsf{U}^{n,1} - \mathsf{U}^{n-1}.\end{aligned}$$

The intermediate vector $\mathsf{U}^{n,1}$ can be eliminated, and the cPG(1) scheme takes again the same form as the Crank–Nicolson scheme, that is,

$$\mathcal{M}(\hat{\mathsf{U}}^n - \mathsf{U}^{n-1}) + \frac{1}{2}\tau\mathcal{A}(\hat{\mathsf{U}}^n + \mathsf{U}^{n-1}) = \tau\mathsf{F}^{n,1}.$$

For $k := 2$, we have $\xi_1 := -\frac{\sqrt{3}}{3}$, $\xi_2 := \frac{\sqrt{3}}{3}$, $\omega_1 := 1$, $\omega_2 := 1$, and the basis functions are

$$\begin{aligned}&\phi_0(s) = L_2(s) = \tfrac{1}{2}(3s^2-1),\\ &\phi_1(s) = -\tfrac{(\sqrt{3}+1)}{4}(s+1)(3s-\sqrt{3}), \quad \psi_1(s) = -\tfrac{1}{2\sqrt{3}}(3s-\sqrt{3}),\\ &\phi_2(s) = \tfrac{(\sqrt{3}-1)}{4}(s+1)(3s+\sqrt{3}), \qquad \psi_2(s) = \tfrac{1}{2\sqrt{3}}(3s+\sqrt{3}).\end{aligned}$$

We obtain

$$\mathbb{B} = \begin{pmatrix} \frac{3}{2} & \frac{2\sqrt{3}-3}{2} \\ -\frac{2\sqrt{3}+3}{2} & \frac{3}{2}\end{pmatrix}, \qquad \mathbb{M} = \begin{pmatrix}\frac{1}{2} & 0 \\ 0 & \frac{1}{2}\end{pmatrix},$$

and

$$d_1 = \sqrt{3}, \quad d_2 = -\sqrt{3}, \qquad \alpha_0 = 1, \quad \alpha_1 = -\sqrt{3}, \quad \alpha_2 = \sqrt{3}.$$

This leads to

$$\begin{aligned}&\frac{3}{2}\mathcal{M}\mathsf{U}^{n,1} + \frac{1}{2}\tau\mathcal{A}\mathsf{U}^{n,1} + \frac{2\sqrt{3}-3}{2}\mathcal{M}\mathsf{U}^{n,2} = \sqrt{3}\mathcal{M}\mathsf{U}^{n-1} + \frac{1}{2}\tau\mathsf{F}^{n,1},\\ &-\frac{2\sqrt{3}+3}{2}\mathcal{M}\mathsf{U}^{n,1} + \frac{3}{2}\mathcal{M}\mathsf{U}^{n,2} + \frac{1}{2}\tau\mathcal{A}\mathsf{U}^{n,2} = -\sqrt{3}\mathcal{M}\mathsf{U}^{n-1} + \frac{1}{2}\tau\mathsf{F}^{n,2},\\ &\mathsf{U}^n = \mathsf{U}^{n-1} - \sqrt{3}(\mathsf{U}^{n,1} - \mathsf{U}^{n,2}).\end{aligned}$$

Notice that the formulae obtained in Steps (i) and (ii) are equivalent.

Exercise 70.5 (Π_τ^k and Π_h commute). Let $v \in H^1(J;V)$. Observe that $\Pi_h(v) \in H^1(J;V)$ so that $\Pi_\tau^k(\Pi_h(v))$ is well defined, and that $\Pi_\tau^k(v) \in H^1(J;V)$ so that $\Pi_h(\Pi_\tau^k(v))$ is well defined as well. Using Remark 70.10, we have

$$\begin{aligned}\Pi_\tau^k(\Pi_h(v)) &= \Pi_h(v)(0) + \int_0^t \Xi_{k-1}^{\mathrm{b}}(\partial_t(\Pi_h(v)))\,\mathrm{d}s\\ &= \Pi_h(v(0)) + \int_0^t \Xi_{k-1}^{\mathrm{b}}(\Pi_h(\partial_t v))\,\mathrm{d}s,\end{aligned}$$

since Π_h commutes with the time derivative. Let us now show that Π_h also commutes with Ξ_{k-1}^{b}. We first observe that

$$\int_J \langle \phi, \Xi_{k-1}^{\mathrm{b}}(w) - w\rangle_{V',V}\mathrm{d}t = 0,$$

for all $\phi \in P^{\mathrm{b}}_{k-1}(\overline{J}_\tau; V')$ and all $w \in L^2(J;V)$ as a consequence of (69.2). This identity, in turn, implies that

$$\begin{aligned}\int_J (\Xi^{\mathrm{b}}_{k-1}(\Pi_h(v)), q_\tau)_L \mathrm{d}t &= \int_J (\Pi_h(v), q_\tau)_L \mathrm{d}t = \int_J (\Pi_h(v), \mathcal{P}_{V_h}(q_\tau))_L \mathrm{d}t \\ &= \int_J \langle v, (\Pi_h)^*(\mathcal{P}_{V_h}(q_\tau))\rangle_{V,V'} \mathrm{d}t = \int_J \langle \Xi^{\mathrm{b}}_{k-1}(v), (\Pi_h)^*(\mathcal{P}_{V_h}(q_\tau))\rangle_{V,V'} \mathrm{d}t \\ &= \int_J (\Pi_h(\Xi^{\mathrm{b}}_{k-1}(v)), \mathcal{P}_{V_h}(q_\tau))_L \mathrm{d}t = \int_J (\Pi_h(\Xi^{\mathrm{b}}_{k-1}(v)), q_\tau)_L \mathrm{d}t,\end{aligned}$$

for all $q_\tau \in P^{\mathrm{b}}_{k-1}(\overline{J}_\tau; L)$, where we used that $\mathcal{P}_{V_h}(q_\tau) \in P^{\mathrm{b}}_{k-1}(\overline{J}_\tau; V_h)$ and $(\Pi_h)^*(\mathcal{P}_{V_h}(q_\tau)) \in P^{\mathrm{b}}_{k-1}(\overline{J}_\tau; V')$. Finally, the above commuting property implies that

$$\Pi^k_\tau(\Pi_h(v)) = \Pi_h(v(0)) + \int_0^t \Pi_h(\Xi^{\mathrm{b}}_{k-1}(\partial_t v))\,\mathrm{d}s = \Pi_h(\Xi^{\mathrm{b}}_{k-1}(v)).$$

This proves the assertion.

Chapter 71

Analysis using inf-sup stability

Exercises

Exercise 71.1 (Time derivative). Let $\phi \in C_0^\infty(J;\mathbb{R})$ and $v \in X$, i.e., $v \in L^2(J;V)$ and $\partial_t v \in L^2(J;V')$. Show that ϕv is in X with $\partial_t(\phi v)(t) = \phi'(t)v(t) + \phi(t)\partial_t v(t)$. (*Hint*: use Pettis theorem and Lemma 64.33.)

Exercise 71.2 (Inf-sup condition). Prove (71.7) with X equipped with the norm $\|v\|_{\tilde{X}}^2 := \|v\|_{L^2(J;V)}^2 + \frac{\gamma}{\alpha}\|\partial_t v\|_{L^2(J;V')}^2 + \gamma\|v(0)\|_L^2$. (*Hint*: use integration by parts in time to bound $\gamma\|v(0)\|_L^2$ by $\|v\|_X^2$.)

Exercise 71.3 (Heat equation). Consider the heat equation with unit diffusivity (see Example 71.4). Prove that for all $v \in X$,

$$\|v\|_X^2 = \sup_{y_1 \in L^2(J;H_0^1)} \frac{b(v,(0,y_1))^2}{\|y_1\|_{L^2(J;H_0^1)}^2} + \|v(0)\|_{L^2}^2.$$

(*Hint*: observe that the supremum is reached for $y_1 := A^{-1}(\partial_t v) + v$.)

Exercise 71.4 (Ultraweak formulation). Equip the space X_{uw} with the norm $\|v\|_{X_{\mathrm{uw}}} := \|v\|_{L^2(J;V)}$ and the space Y_{uw} with the norm defined in (71.10). (i) Prove the inf-sup condition (71.11). (*Hint*: consider the adjoint parabolic problem $\partial w_v(t) + A^*(w_v)(t) := (v(t),\cdot)_V$ for a.e. $t \in J$, with $w_v(0) := 0$, invoke Lemma 71.2, then set $\tilde{w}_v(t) := w_v(T-t)$.) (ii) The rest of the exercise considers the heat equation with unit diffusivity. Show that $\sup_{w\in Y_{\mathrm{uw}}} \frac{b_{\mathrm{uw}}(v,w)}{\|w\|_{Y_{\mathrm{uw}}}} \le \|v\|_{X_{\mathrm{uw}}}$ for all $v \in X_{\mathrm{uw}}$. (*Hint*: prove first that $\|A^{-1}(\partial_t w) - w\|_{L^2(J;H_0^1(D))}^2 = \|w\|_{Y_{\mathrm{uw}}}^2$ for all $w \in Y_{\mathrm{uw}}$.) (iii) Prove that

$$\|v\|_{X_{\mathrm{uw}}} = \sup_{w\in Y_{\mathrm{uw}}} \frac{b_{\mathrm{uw}}(v,w)}{\|w\|_{Y_{\mathrm{uw}}}}, \qquad \forall v \in X_{\mathrm{uw}}.$$

(*Hint*: compute $b(v,\widetilde{w})$, where $\widetilde{w}_v \in Y_{\mathrm{uw}}$ solve the backward-in-time parabolic problem $-\partial_t \widetilde{w}_v - \Delta\widetilde{w}_v = -\Delta v$ with $\widetilde{w}_v(T) = 0$.)

Exercise 71.5 (Norm $\|\cdot\|_{V_h'}$). Let $\|\cdot\|_{V_h'}$ be defined in (71.13). Let $\{\varphi_i\}_{i\in\{1:I\}}$ be a basis of V_h and let $\mathcal{S} \in \mathbb{R}^{I\times I}$ and $\mathcal{M} \in \mathbb{R}^{I\times I}$ be the stiffness and mass matrices s.t. $\mathcal{S}_{ij} := (\varphi_j,\varphi_i)_V$ and $\mathcal{M}_{ij} := (\varphi_j,\varphi_i)_L$ for all $i,j \in \{1{:}I\}$ (these matrices are symmetric positive definite). For all

$v_h \in V_h$, let $\mathsf{V} \in \mathbb{R}^I$ be the coordinate vector of v_h in the basis $\{\varphi_i\}_{i\in\{1:I\}}$, i.e., $v_h := \sum_{i\in\{1:I\}} \mathsf{V}_i\varphi_i$. (i) Prove that $\|v_h\|_{V_h'} = (\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V})^{\frac12}$. (*Hint*: use that $\|v_h\|_{V_h'} = \sup_{\mathsf{W}\in\mathbb{R}^I} \frac{\mathsf{W}^\mathsf{T}\mathcal{M}\mathsf{V}}{(\mathsf{W}^\mathsf{T}\mathcal{S}\mathsf{W})^{\frac12}}$.) (ii) Let $\mu \ge 0$. Prove the following two-sided bound due to Pearson and Wathen [37] (see also Smears [41]):

$$\frac12 \le \frac{\mathsf{V}^\mathsf{T}(\mathcal{M}\mathcal{S}^{-1}\mathcal{M}+\mu\mathcal{S})\mathsf{V}}{\mathsf{V}^\mathsf{T}(\mathcal{M}+\mu^{\frac12}\mathcal{S})\mathcal{S}^{-1}(\mathcal{M}+\mu^{\frac12}\mathcal{S})\mathsf{V}} \le 2, \qquad \forall \mathsf{V}\in\mathbb{R}^I.$$

Exercise 71.6 (Error analysis with $\|\cdot\|_{X_h}$). Referring to §71.2 and denoting by u_h the solution to (71.12), let $\eta(t) := u(t) - \mathcal{P}_{V_h}(u(t))$ for a.e. $t \in J$. (i) With the norm $\|\cdot\|_{X_h}$ defined in (71.15), prove that $|b(\eta, y_h)| \le \sqrt{2}M\|\eta\|_{X_h}\|y_h\|_Y$ for all $y_h \in Y_h$. (*Hint*: use that $\frac{\alpha}{\gamma_h} \le M^2$.) (ii) Prove the error estimate $\|u - u_h\|_{X_h} \le \Big(1 + \frac{\sqrt{2}M}{\beta_h}\Big)\|\eta\|_{X_h}$, where β_h is the constant from the inf-sup inequality (71.16). (*Hint*: combine inf-sup stability with consistency and boundedness.)

Exercise 71.7 ($C^0(\overline{J};L)$-estimate using inf-sup stability). (i) Recalling that $\|\cdot\|_X$ is defined in (71.6a), prove that $\gamma^{\frac12}\|v\|_{C^0(\overline{J};L^2)} \le \|v\|_X$ for all $v \in X$. (*Hint*: see Exercise 71.2.) (ii) Assume (71.18). Let $c_1 := \sqrt{\frac{\gamma}{\alpha}}$ and $c_2 := \sqrt{\frac{\rho}{2}}$, where $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$ and $\iota_{L,V}$ is the operator norm of the embedding $V \hookrightarrow L$, i.e., the smallest constant s.t. $\|v\|_L \le \iota_{L,V}\|v\|_V$ for all $v \in V$. Prove that

$$\beta_h' c_1\|u - u_h\|_{C^0(\overline{J};L)} \le \beta_h' c_1\|\eta\|_{C^0(\overline{J};L)} + \|\eta(0)\|_L + c_2\|\partial_t\eta\|_{L^2(J;L)},$$

with $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(t;u(t))$. (*Hint*: combine Lemma 71.9 with consistency.) (iii) Compare this estimate with (66.16) in the context of the heat equation.

Exercise 71.8 (Implicit Euler scheme). (i) Let $X_{h\tau} := (V_h)^{N+1}$ and $Y_{h\tau} := V_h\times(V_h)^N$. Reformulate the implicit Euler scheme (67.3) using the forms

$$\begin{aligned} b_\tau(v_{h\tau}, y_{h\tau}) &:= (v_h^0, y_{0h})_L + \sum_{n\in\mathcal{N}_\tau} \tau\big(((\delta_\tau v_{h\tau})^n, y_{1h}^n)_L + a^n(v_h^n, y_{1h}^n)\big),\\ \ell_\tau(y_{h\tau}) &:= (u_0, y_{0h})_L + \sum_{n\in\mathcal{N}_\tau} \tau\langle f^n, y_{1h}^n\rangle_{V',V}, \end{aligned}$$

where $(\delta_\tau v_{h\tau})^n := \frac{1}{\tau}(v_h^n - v_h^{n-1})$. (ii) Assume that the bilinear form a is symmetric at all times. Prove that

$$\alpha\|u_{h\tau}\|_{\ell^2(J;V)}^2 + \frac{1}{M}\|\delta_\tau u_{h\tau}\|_{\ell^2(J;V_h')}^2 + \tau\|\delta_\tau u_{h\tau}\|_{\ell^2(J;L)}^2 + \|u_h^N\|_L^2 \le \frac{M}{\alpha}\Big(\frac{1}{\alpha}\|f\|_{\ell^2(J;V')}^2 + \|u_0\|_L^2\Big).$$

(*Hint*: use the inf-sup condition (67.1).) (iii) Assume that $u \in C^0(\overline{J};V)\cap C^1(\overline{J};V')\cap H^2(J;V')$ and that $\mathcal{P}_{V_h}$ is uniformly V-stable (see (71.18)). Prove that

$$\begin{aligned} \|\delta_\tau u_{h\tau} - \delta_\tau u_\tau\|_{\ell^2(J;V')} \le \|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)}\frac{M}{\alpha}\Bigg(&\sqrt{3}\big(M\|\eta_\tau\|_{\ell^2(J;V)} + 2\|\partial_t\eta\|_{L^2(J;V')}\\ &+ \tau\|\partial_{tt}u\|_{L^2(J;V')}\big) + \sqrt{\alpha}\|e_h^0\|_L\Bigg), \end{aligned}$$

where $(\delta_\tau u_\tau)^n := \frac{1}{\tau}(u(t_n) - u(t_{n-1}))$ for all $n \in \mathcal{N}_\tau$, $\eta(t) := u(t) - v_h(t)$ for all $t \in \overline{J}$, $\eta_\tau := (\eta(t_n))_{n\in\mathcal{N}_\tau}$, and v_h arbitrary in $H^1(J;V_h)$. (*Hint*: use Step (ii) and Lemma 71.8.)

Exercise 71.9 (Inf-sup for cPG(k)). Complete the proof of Lemma 71.20. (*Hint*: reason as in the last step of the proof of Lemma 71.18.)

Solution to exercises

Exercise 71.1 (Time derivative). Let $\phi \in C_0^\infty(J;\mathbb{R})$ and $v \in X$. The function $J \ni t \mapsto \phi(t)v(t) \in V$ is strongly measurable owing to Pettis theorem (Theorem 64.4). Indeed V is a separable Hilbert space (by assumption), and the function $J \ni t \mapsto \phi(t)\langle v', v(t)\rangle_{V',V} \in \mathbb{R}$ is Lebesgue measurable for all $v' \in V'$. Moreover $(\int_J \|\phi(t)v(t)\|_V^2 \mathrm{d}t)^{\frac12} \le \|\phi\|_{L^\infty(J;\mathbb{R})}\|v\|_{L^2(J;V)}$. This shows that $\phi v \in L^2(J;V)$ (see Definition 64.17).
Let now $\psi \in C_0^\infty(J;\mathbb{R})$ and $w \in V$. Since $\phi(t)\psi'(t) = (\phi\psi)'(t) - \phi'(t)\psi(t)$ for all $t \in J$, we have

$$\begin{aligned}\int_J (w, \phi(t)v(t))_L \psi'(t)\mathrm{d}t &= \int_J (w, v(t))_L (\phi\psi)'(t)\mathrm{d}t - \int_J (w, v(t))_L \phi'(t)\psi(t)\mathrm{d}t \\ &= \int_J -\langle \partial_t v(t), w\rangle_{V',V}\phi(t)\psi(t)\mathrm{d}t - \int_J (w, v(t))_L \phi'(t)\psi(t)\mathrm{d}t \\ &= \int_J -\langle \phi(t)\partial_t v(t), w\rangle_{V',V}\psi(t)\mathrm{d}t - \int_J (\phi'(t)v(t), w)_L \psi(t)\mathrm{d}t,\end{aligned}$$

where we used that v has a weak time derivative in $L^2(J;V')$ and that the L-inner product is an extension of the duality between V' and V. The above identity together with the characterization from Lemma 64.33 shows that ϕv has a weak time derivative in $L^2(J;V')$ such that $\partial_t(\phi v)(t) = \phi'(t)v(t) + \phi(t)\partial_t v(t)$.

Exercise 71.2 (Inf-sup condition). Let $v \in X$. Integrating by parts in time and using Young's inequality, we infer that

$$\begin{aligned}\gamma\|v(0)\|_L^2 &= \gamma\|v(T)\|_L^2 - 2\gamma\int_J \langle \partial_t v(t), v(t)\rangle_{V',V}\mathrm{d}t \\ &\le \gamma\|v(T)\|_L^2 + \frac{\gamma}{\alpha}\|\partial_t v\|_{L^2(J;V')}^2 + \gamma\alpha\|v\|_{L^2(J;V)}^2.\end{aligned}$$

Since $\gamma\alpha \le 1$, this implies that

$$\gamma\|v(0)\|_L^2 \le \frac{1}{\alpha}\|v(T)\|_L^2 + \frac{\gamma}{\alpha}\|\partial_t v\|_{L^2(J;V')}^2 + \|v\|_{L^2(J;V)}^2 = \|v\|_X^2.$$

Hence, $\|v\|_{\tilde X}^2 \le 2\|v\|_X^2$. We conclude that

$$\inf_{v\in X}\sup_{y\in Y}\frac{|b(v,y)|}{\|v\|_{\tilde X}\|y\|_Y} \ge \tilde\beta > 0,$$

with $\tilde\beta := \frac{1}{\sqrt2}\beta$, and β is the inf-sup constant in (71.8).

Exercise 71.3 (Heat equation). Let $v \in X := L^2(J;H_0^1(D))$. We have

$$\begin{aligned}b(v,(0,y_1)) &= \int_J \langle \partial_t v(t) + A(v(t)), y_1(t)\rangle_{H^{-1}(D),H_0^1(D)}\mathrm{d}t \\ &= \int_J \langle A(A^{-1}\partial_t v(t) + v(t)), y_1(t)\rangle_{H^{-1}(D),H_0^1(D)}\mathrm{d}t \\ &= \int_J (\nabla(A^{-1}(\partial_t v(t)) + v(t)), \nabla y_1(t))_{\boldsymbol{L}^2(D)}\mathrm{d}t.\end{aligned}$$

Letting $y_v := A^{-1}(\partial_t v) + v \in Y_1 := L^2(J;H_0^1(D))$, this shows that

$$b(v,(0,y_1)) = \int_0^T (\nabla y_v, \nabla y_1)_{\boldsymbol{L}^2(D)}\mathrm{d}t.$$

Hence, as claimed in the hint, we have

$$\sup_{y_1\in L^2(J;H^1_0(D))}\frac{b(v,(0,y_1))}{\|y_1\|_{L^2(J;H^1_0(D))}}=\sup_{y_1\in L^2(J;H^1_0(D))}\frac{\int_J(\nabla y_v(t),\nabla y_1(t))_{\boldsymbol{L}^2(D)}\mathrm{d}t}{\|y_1\|_{L^2(J;H^1_0(D))}}$$
$$=\|y_v\|_{L^2(J;H^1_0(D))}.$$

This gives

$$b(v,(0,y_v))=\|y_v\|^2_{L^2(J;H^1_0(D))}=\left(\sup_{y_1\in L^2(J;H^1_0(D))}\frac{b(v,(0,y_1))}{\|y_1\|_{L^2(J;H^1_0(D))}}\right)^2.$$

Moreover, since A is self-adjoint and observing that

$$\langle A(y),z\rangle_{H^{-1}(D),H^1_0(D)}=(\nabla y,\nabla z)_{\boldsymbol{L}^2(D)}=(y,z)_{H^1_0(D)},\qquad\forall y,z\in H^1_0(D),$$

i.e., A is the Riesz–Fréchet isometry from $H^1_0(D)$ to $H^{-1}(D)$, we have

$$\begin{aligned}b(v,(0,y_v))&=\int_J\langle\partial_t v(t)+A(v(t)),A^{-1}(\partial_t v(t))+v(t)\rangle_{H^{-1}(D),H^1_0(D)}\mathrm{d}t\\&=2\int_J\langle\partial_t v(t),v(t)\rangle_{H^{-1}(D),H^1_0(D)}\mathrm{d}t+\int_J\langle A(v(t)),v(t)\rangle_{H^{-1}(D),H^1_0(D)}\mathrm{d}t\\&\quad+\int_J\langle\partial_t v(t),A^{-1}(\partial_t v(t))\rangle_{H^{-1}(D),H^1_0(D)}\mathrm{d}t\\&=\|v(T)\|^2_{L^2(D)}-\|v(0)\|^2_{L^2(D)}+\|v\|^2_{L^2(J;H^1_0(D))}+\|\partial_t v\|^2_{L^2(J;H^{-1}(D))}\\&=\|v\|^2_X-\|v(0)\|^2_{L^2(D)},\end{aligned}$$

where we used integration by parts in time. This proves the inf-sup identity.

Exercise 71.4 (Ultraweak formulation). (i) Let $v\in X_{\mathrm{uw}}$. Consider the adjoint parabolic problem $\partial w_v(t)+A^*(w_v)(t)=(v(t),\cdot)_V$ for a.e. $t\in J$, with the initial condition $w_v(0)=0$ (this problem is well-posed owing to Theorem 65.9). Since the operators A and A^{-1} have the same coercivity constants as A^* and A^{-*}, respectively, and since $w_v(0)=0$, Lemma 71.2 implies that

$$\begin{aligned}\beta\,\|w_v\|_X&\le\sup_{y_1\in L^2(J;V)}\frac{\int_J\langle\partial_t w_v(t)+A^*(w_v)(t),y_1(t)\rangle_{V',V}\mathrm{d}t}{\|y_1\|_{L^2(J;V)}}\\&=\sup_{y_1\in L^2(J;V)}\frac{\int_J(v(t),y_1(t))_V\mathrm{d}t}{\|y_1\|_{L^2(J;V)}}=\|v\|_{L^2(J;V)}.\end{aligned}$$

Let us now set

$$\widetilde{w}_v(t):=w_v(T-t).$$

Then $\widetilde{w}_v\in Y_{\mathrm{uw}}$ and since $\partial_t\widetilde{w}_v(t)=-\partial_t w_v(T-t)$, we infer that $-\partial_t\widetilde{w}_v(t)+A^*(\widetilde{w}_v)(t)=(v(t),\cdot)_V$, for a.e. $t\in J$. This implies that $b_{\mathrm{uw}}(v,\widetilde{w}_v)=\|v\|^2_{L^2(J;V)}$. Since $\|\widetilde{w}_v\|_{Y_{\mathrm{uw}}}=\|w_v\|_X\le\beta^{-1}\|v\|_{L^2(J;V)}$, we conclude that

$$\|v\|_{X_{\mathrm{uw}}}=\|v\|_{L^2(J;V)}=\frac{b_{\mathrm{uw}}(v,\widetilde{w}_v)}{\|v\|_{L^2(J;V)}}\le\beta^{-1}\frac{b_{\mathrm{uw}}(v,\widetilde{w}_v)}{\|\widetilde{w}_v\|_{Y_{\mathrm{uw}}}}\le\beta^{-1}\sup_{w\in Y_{\mathrm{uw}}}\frac{b_{\mathrm{uw}}(v,w)}{\|w\|_{Y_{\mathrm{uw}}}}.$$

(ii) Let $w \in Y_{\rm uw}$. Using that $\langle g, z\rangle_{H^{-1}(D),H^1_0(D)} = (\nabla A^{-1}(g), \nabla z)_{\boldsymbol{L}^2}$ for all $g \in H^{-1}(D)$ and all $z \in H^1_0(D)$, we infer that

$$\begin{aligned}\|A^{-1}(\partial_t w) - w\|^2_{L^2(J;H^1_0(D))} &= \int_J \langle \partial_t w(t) - A(w(t)), A^{-1}(\partial_t w(t)) - w(t)\rangle_{H^{-1}(D),H^1_0(D)}\mathrm{d}t\\ &= \|w\|^2_{L^2(J;H^1_0(D))} + \|\partial_t w\|^2_{L^2(J;H^{-1}(D))} + \|w(0)\|^2_{L^2(D)} = \|w\|^2_{Y_{\rm uw}}.\end{aligned}$$

This shows that $\|A^{-1}(\partial_t w) - w\|_{L^2(J;H^1_0(D))} = \|w\|_{Y_{\rm uw}}$ for all $w \in Y_{\rm uw}$. Let now $v \in X_{\rm uw} := L^2(J; H^1_0(D))$. Proceeding as above, we have

$$\begin{aligned}b_{\rm uw}(v, w) &= \int_J \langle v(t), -\partial_t w(t)\rangle_{H^1_0(D),H^{-1}(D)} + (\nabla v(t), \nabla w(t))_{\boldsymbol{L}^2(D)}\mathrm{d}t\\ &= \int_J (\nabla v(t), \nabla(-A^{-1}(\partial_t w(t)) + w(t))_{\boldsymbol{L}^2(D)}\mathrm{d}t.\end{aligned}$$

This implies that for all $v \in X_{\rm uw}$,

$$\sup_{w\in Y_{\rm uw}} \frac{b_{\rm uw}(v,w)}{\|w\|_{Y_{\rm uw}}} \le \|v\|_{L^2(J;H^1_0(D))} \sup_{w\in Y_{\rm uw}} \frac{\| - A^{-1}(\partial_t w(t)) + w(t)\|_{L^2(J;H^1_0(D))}}{\|w\|_{Y_{\rm uw}}}.$$

Hence, $\sup_{w\in Y_{\rm uw}} \frac{b_{\rm uw}(v,w)}{\|w\|_{Y_{\rm uw}}} \le \|v\|_{L^2(J;H^1_0(D))}$.

(iii) Let $v \in X_{\rm uw} := L^2(J; H^1_0(D))$, $v \neq 0$. Let $\widetilde{w}_v \in Y_{\rm uw}$ solve the backward-in-time parabolic problem $-\partial_t\widetilde{w}_v - \Delta\widetilde{w}_v := -\Delta v$ in $D\times J$ with homogeneous Dirichlet boundary conditions and the final condition $\widetilde{w}_v(T) := 0$. Setting $w_v(t) := \widetilde{w}_v(T-t)$, we observe that $w_v \in X$ and that w_v satisfies the heat equation $\partial_t w_v(t) - \Delta w_v(t) = -\Delta v(T-t)$ with the initial condition $w_v(0) = 0$. Defining $f \in L^2(J; H^{-1}(D))$ s.t. $f(t) := -\Delta v(T-t)$ for a.e. $t \in J$, we have

$$\int_J \big(\langle \partial_t w_v(t), \theta(t)\rangle_{H^{-1}(D),H^1_0(D)} + (\nabla w_v(t), \nabla\theta(t))_{\boldsymbol{L}^2(D)}\big)\mathrm{d}t = \int_J \langle f(t), \theta(t)\rangle_{H^{-1}(D),H^1_0(D)}\mathrm{d}t,$$

for all $\theta \in L^2(J; H^1_0(D))$. Theorem 65.9 shows that the problem defining w_v is well-posed. Hence, the function $\widetilde{w}_v \in Y_{\rm uw}$ is well defined. Since $A(v) = -\partial_t\widetilde{w}_v + A(\widetilde{w})$, i.e., $v := -A^{-1}(\partial_t\widetilde{w}_v) + \widetilde{w}$, we infer from Step (ii) that $b_{\rm uw}(v, \widetilde{w}_v) = \|v\|^2_{L^2(J;H^1_0(D))} = \|A^{-1}(\partial_t(\widetilde{w}_v)) - \widetilde{w}_v\|^2_{L^2(J;H^1_0(D))}$. Therefore, we have

$$\|v\|_{L^2(J;H^1_0(D))} \ge \sup_{w\in Y_{\rm uw}} \frac{b_{\rm uw}(v,w)}{\|w\|_{Y_{\rm uw}}} \ge \frac{b_{\rm uw}(v,\widetilde{w}_v)}{\|\widetilde{w}_v\|_{Y_{\rm uw}}} = \|v\|_{L^2(J;H^1_0(D))}.$$

This proves the inf-sup identity.

Exercise 71.5 (Norm $\|\cdot\|_{V_h'}$). Note that both matrices $\mathcal{S}$ and $\mathcal{M}$ are symmetric positive definite.
(i) Since $\mathcal{M}$ and $\mathcal{S}$ are both invertible and symmetric, using the hint we infer that

$$\|v_h\|_{V_h'} = \sup_{\mathsf{Z}\in\mathbb{R}^I} \frac{\mathsf{Z}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V}}{(\mathsf{Z}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{S}\mathcal{S}^{-1}\mathcal{M}\mathsf{Z})^{\frac12}} = \sup_{\mathsf{Z}\in\mathbb{R}^I} \frac{\mathsf{Z}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V}}{(\mathsf{Z}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{Z})^{\frac12}}.$$

But the Cauchy–Schwarz inequality implies that

$$(\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V})^{\frac12} \le \sup_{\mathsf{Z}\in\mathbb{R}^I} \frac{\mathsf{Z}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V}}{(\mathsf{Z}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{Z})^{\frac12}} \le (\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V})^{\frac12},$$

whence the result.
Note: we can also prove the result using a Lagrange multiplier technique as presented in §49.3.1.

Using the hint, we infer that $\|v_h\|_{V_h'} = \mathsf{V}^\mathsf{T}\mathcal{M}\mathsf{W}_*$, where $\mathsf{W}_* \in \mathbb{R}^I$ solves the following linear maximization problem under a quadratic constraint:

$$\mathsf{W}_* = \arg\max_{\substack{\mathsf{W}\in\mathbb{R}^I \\ \mathsf{W}^\mathsf{T}\mathcal{S}\mathsf{W}=1}} \mathsf{W}^\mathsf{T}\mathcal{M}\mathsf{V}.$$

The optimality conditions characterizing the unique solution of the above problem can be formulated by introducing a Lagrange multiplier λ and the Lagrange functional $\mathcal{L}(\mathsf{W},\lambda) := \mathsf{W}^\mathsf{T}\mathcal{M}\mathsf{V} + \frac{1}{2}\lambda(\mathsf{W}^\mathsf{T}\mathcal{S}\mathsf{W}-1)$. The pair (W_*,λ_*) is extremal for $\mathcal{L}$ iff (see §49.3.1):

$$\mathcal{M}\mathsf{V} + \lambda_*\mathcal{S}\mathsf{W}_* = 0, \qquad \mathsf{W}_*^\mathsf{T}\mathcal{S}\mathsf{W}_* = 1.$$

This implies that $\mathsf{W}_* = -\lambda_*\mathcal{S}^{-1}\mathcal{M}\mathsf{V}$, and inserting this expression into the constraint $\mathsf{W}_*^\mathsf{T}\mathcal{S}\mathsf{W}_* = 1$ and using the symmetry of $\mathcal{S}$ and $\mathcal{M}$ implies that $\lambda_* = \pm(\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V})^{-\frac{1}{2}}$. We conclude that

$$\mathsf{W}_*^\mathsf{T}\mathcal{M}\mathsf{V} = \mp(\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V})^{-\frac{1}{2}}\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V} = \mp(\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V})^{\frac{1}{2}},$$

which, in turn, implies that λ_* must be negative. This completes the proof.
(ii) Let $\mu \geq 0$. The upper bond follows from the identity

$$(\mathcal{M}+\mu^{\frac{1}{2}}\mathcal{S})\mathcal{S}^{-1}(\mathcal{M}+\mu^{\frac{1}{2}}\mathcal{S}) = \mathcal{M}\mathcal{S}^{-1}\mathcal{M} + 2\mu^{\frac{1}{2}}\mathcal{M} + \mu\mathcal{S},$$

and the fact that $\mathcal{M}$ is positive definite. To prove the lower bound, we first infer from Step (i) that for all $\mathsf{V}\in\mathbb{R}^I$,

$$(\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V})^{\frac{1}{2}} = \|v_h\|_{V_h'} = \sup_{\mathsf{W}\in\mathbb{R}^I} \frac{\mathsf{V}^\mathsf{T}\mathcal{M}\mathsf{W}}{(\mathsf{W}^\mathsf{T}\mathcal{S}\mathsf{W})^{\frac{1}{2}}} \geq \frac{\mathsf{V}^\mathsf{T}\mathcal{M}\mathsf{V}}{(\mathsf{V}^\mathsf{T}\mathcal{S}\mathsf{V})^{\frac{1}{2}}},$$

where the last bound follows by taking $\mathsf{W} := \mathsf{V}$. Hence, we have

$$\mathsf{V}^\mathsf{T}\mathcal{M}\mathsf{V} \leq (\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V})^{\frac{1}{2}}(\mathsf{V}^\mathsf{T}\mathcal{S}\mathsf{V})^{\frac{1}{2}}.$$

Recalling that $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ for every real numbers $a,b\in\mathbb{R}$, we infer that

$$\begin{aligned}\mathsf{V}^\mathsf{T}(\mathcal{M}+\mu^{\frac{1}{2}}\mathcal{S})\mathcal{S}^{-1}(\mathcal{M}+\mu^{\frac{1}{2}}\mathcal{S})\mathsf{V} &= \mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V} + 2\mu^{\frac{1}{2}}\mathsf{V}^\mathsf{T}\mathcal{M}\mathsf{V} + \mu\mathsf{V}^\mathsf{T}\mathcal{S}\mathsf{V} \\ &\leq 2\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V} + 2\mu\mathsf{V}^\mathsf{T}\mathcal{S}\mathsf{V},\end{aligned}$$

which proves the lower bound.

Exercise 71.6 (Error analysis with $\|\cdot\|_{X_h}$). (i) Let $y_h \in Y_h$. We have

$$\begin{aligned}b(\eta,y_h) &= (\eta(0),y_{0h})_L + \int_J \langle\partial_t\eta(t) + A(\eta)(t), y_{1h}(t)\rangle_{V',V}\mathrm{d}t \\ &= \int_J \langle\partial_t\eta(t) + A(\eta)(t), y_{1h}(t)\rangle_{V',V}\mathrm{d}t \\ &\leq \big(\|\partial_t\eta\|_{L^2(J;V_h')} + M\|\eta\|_{L^2(J;V)}\big)\|y_{1h}\|_{L^2(J;V)} \\ &\leq \sqrt{2}\big(\|\partial_t\eta\|_{L^2(J;V_h')}^2 + M^2\|\eta\|_{L^2(J;V)}^2\big)^{\frac{1}{2}}\|y_h\|_Y \\ &\leq \sqrt{2}M\|\eta\|_{X_h}\|y_h\|_Y,\end{aligned}$$

where we used that $(\eta(0),y_{0h})_L = 0$ in the second line and that $\frac{1}{M^2} \leq \frac{\gamma_h}{\alpha}$ in the last line. This proves the assertion. Notice in passing that it is not possible to prove the uniform boundedness of

b on $X\times Y_h$ if X is equipped with the $\|\cdot\|_{X_h}$-norm unless the initial value of the first argument is L-orthogonal to V_h.
(ii) Let us set $e_h(t) := u_h(t) - \mathcal{P}_{V_h}(u(t))$ for a.e. $t\in J$. Combining the inf-sup inequality (71.16) with consistency (Galerkin orthogonality) and boundedness (Step (i)), we infer that

$$\beta\|e_h\|_{X_h} \le \sup_{y_h\in Y_h}\frac{b(e_h,y_h)}{\|y_h\|_Y} = \sup_{y_h\in Y_h}\frac{b(\eta,y_h)}{\|y_h\|_Y} \le \sqrt{2}M\|\eta\|_{X_h},$$

and the error estimate follows from the triangle inequality.

Exercise 71.7 ($C^0(\overline{J};L)$-estimate using inf-sup stability). (i) It suffices to repeat the argument from Exercise 71.2 by replacing the time integration over J by the time integration over (t,T) for all $t\in[0,T)$. We obtain

$$\begin{aligned}\gamma\|v(t)\|_L^2 &= \gamma\|v(T)\|_L^2 - 2\gamma\int_{(t,T)}\langle\partial_t v(t),v(t)\rangle_{V',V}\mathrm{d}t\\ &\le \gamma\|v(T)\|_L^2 + \frac{\gamma}{\alpha}\|\partial_t v\|_{L^2((t,T);V')}^2 + \gamma\alpha\|v\|_{L^2((t,T);V)}^2.\end{aligned}$$

Since $\gamma\alpha\le 1$ and $(t,T)\subset J$, this implies that

$$\gamma\|v(t)\|_L^2 \le \frac{1}{\alpha}\|v(T)\|_L^2 + \frac{\gamma}{\alpha}\|\partial_t v\|_{L^2(J;V')}^2 + \|v\|_{L^2(J;V)}^2 = \|v\|_X^2,$$

and the claim follows by taking the supremum over $t\in\overline{J}$ on the left-hand side (there is nothing to prove if $t=T$).
(ii) Let us set $e_h(t) := u_h(t) - \Pi_h^{\mathrm{E}}(t;u(t))$ and $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(t;u(t))$ for a.e. $t\in J$ (recall that $\Pi_h^{\mathrm{E}}(t;u(t))$ is the elliptic projection operator defined in (66.11)). Using the inf-sup condition from Lemma 71.9, we infer that

$$\beta_h'\gamma^{\frac12}\|e_h\|_{C^0(\overline{J};L)} \le \beta_h'\|e_h\|_X \le \sup_{y_h\in Y_h}\frac{b(e_h,y_h)}{\|y_h\|_Y}.$$

Consistency gives $b(e_h,y_h) = b(\eta,y_h)$ (this identity is often called Galerkin orthogonality), and the definition of the elliptic projection implies that

$$b(e_h,y_h) = (\eta(0),y_{0h})_L + \int_J\langle\partial_t\eta(t),y_{1h}(t)\rangle_{V',V}\mathrm{d}t.$$

Since $\|y_h\|_Y^2 = \frac{1}{\alpha}\|y_{0h}\|_L^2 + \|y_{1h}\|_{L^2(J;V)}^2$, we obtain

$$\beta_h'\gamma^{\frac12}\|e_h\|_{C^0(\overline{J};L)} \le \alpha^{\frac12}\|\eta(0)\|_L + \|\partial_t\eta\|_{L^2(J;V')}.$$

Denoting by $\iota_{L,V}$ the operator norm of the embedding $V\hookrightarrow L$, i.e., the smallest constant s.t. $\|v\|_L\le\iota_{L,V}\|v\|_V$ for all $v\in V$, we have $\|\partial_t\eta\|_{L^2(J;V')}\le\iota_{L,V}\|\partial_t\eta\|_{L^2(J;L)}$. Dividing by $\alpha^{\frac12}$ and using the triangle inequality, we obtain

$$\beta_h'\frac{\gamma^{\frac12}}{\alpha^{\frac12}}\|u-u_h\|_{C^0(\overline{J};L)} \le \|\eta(0)\|_L + \beta_h'\frac{\gamma^{\frac12}}{\alpha^{\frac12}}\|\eta\|_{C^0(\overline{J};L)} + \frac{\iota_{L,V}}{\alpha^{\frac12}}\|\partial_t\eta\|_{L^2(J;V')}.$$

This proves the assertion with the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$.
(iii) In the context of the heat equation (i.e., $L:=L^2(D)$, $V:=H_0^1(D)$, the operator A is time-independent and self-adjoint), the estimate obtained above is very similar to (66.16). However, (66.16) is sharper for the following three reasons: (1) it includes weights with exponential decay in time; (2) the norm involving $\partial_t\eta$ is only integrated over $(0,t)$ and not over J; (3) the constant $\beta_h'\sqrt{\gamma/\alpha}$ is smaller than 1 (because $\beta_h'\sqrt{\gamma/\alpha}\le\beta\sqrt{\gamma/\alpha}\le\gamma\alpha\le 1$).

Exercise 71.8 (Implicit Euler scheme). (i) Let us set $u_h^0 := \mathcal{P}_{V_h}(u_0)$. It is readily seen that $u_{h\tau} := (u_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$ solves (67.3) if and only if $(u_h^0, u_{h\tau}) =: \tilde{u}_{h\tau} \in X_{h\tau}$ solves

$$b_\tau(\tilde{u}_{h\tau}, y_{h\tau}) = \ell_\tau(y_{h\tau}), \qquad \forall y_{h\tau} \in Y_{h\tau}.$$

(ii) Recall that the norms used the inf-sup condition (67.1) are

$$\|v_{h\tau}\|_{X_{h\tau}}^2 := \frac{1}{\alpha}\|v_h^N\|_L^2 + \|v_{h\tau}\|_{\ell^2(J;V)}^2 + \frac{1}{\alpha M}\|\delta_\tau v_{h\tau}\|_{\ell^2(J;V_h')}^2 + \frac{\tau}{\alpha}\|\delta_\tau v_{h\tau}\|_{\ell^2(J;L)}^2,$$
$$\|y_{h\tau}\|_{Y_{h\tau}}^2 := \frac{1}{\alpha}\|y_{0h}\|_L^2 + \|y_{1h\tau}\|_{\ell^2(J;V)}^2.$$

We infer from Step (i) and the inf-sup condition (67.1) that

$$\|\tilde{u}_{h\tau}\|_{X_{h\tau}}^2 \le \frac{M}{\alpha^3}\sup_{y_{h\tau}\in Y_{h\tau}}\frac{|b_\tau(\tilde{u}_{h\tau}, y_{h\tau})|^2}{\|y_{h\tau}\|_{Y_{h\tau}}^2} \le \frac{M}{\alpha^3}\sup_{y_{h\tau}\in Y_{h\tau}}\frac{|\ell_\tau(y_{h\tau})|^2}{\|y_{h\tau}\|_{Y_{h\tau}}^2},$$

and the Cauchy–Schwarz inequality implies that

$$|\ell_\tau(y_{h\tau})|^2 \le (\|f\|_{\ell^2(J;V')}^2 + \alpha\|u_0\|_L^2)\|y_{h\tau}\|_{Y_{h\tau}}^2.$$

The assertion follows readily.

(iii) Proceeding as in the proof of Theorem 67.6, but using the stability estimate from Step (ii) (instead of the estimate (67.7) from Lemma 67.3), and keeping only the term related to the time derivative measured in $\ell^2(J;V_h')$ on the left-hand side, we infer that

$$\frac{1}{M}\|\delta_\tau e_{h\tau}\|_{\ell^2(J;V_h')}^2 \le \frac{M}{\alpha^2}\|g_\tau\|_{\ell^2(J;V')}^2 + \frac{M}{\alpha}\|e_h^0\|_L^2,$$

where $g_\tau := (g^n)_{n\in\mathcal{N}_\tau} \in (V')^N$ and g^n is defined in the proof of Theorem 67.6. Using the bound on $\|g_\tau\|_{\ell^2(J;V')}^2$ from this proof, we obtain

$$\|\delta_\tau e_{h\tau}\|_{\ell^2(J;V_h')}^2 \le \frac{M^2}{\alpha^2}3\big(M^2\|\eta_\tau\|_{\ell^2(J;V)}^2 + \|\partial_t\eta\|_{L^2(J;V')}^2 + \tau^2\|\partial_{tt}u\|_{L^2(J;V')}^2\big) + \frac{M^2}{\alpha}\|e_h^0\|_L^2.$$

Taking the square root yields

$$\|\delta_\tau e_{h\tau}\|_{\ell^2(J;V_h')} \le \frac{M}{\alpha}\sqrt{3}\big(M\|\eta_\tau\|_{\ell^2(J;V)} + \|\partial_t\eta\|_{L^2(J;V')} + \tau\|\partial_{tt}u\|_{L^2(J;V')}\big) + \frac{M}{\alpha}\sqrt{\alpha}\|e_h^0\|_L.$$

The uniform V-stability of the L-orthogonal projection (see Lemma 71.8) gives

$$\|\delta_\tau e_{h\tau}\|_{\ell^2(J;V')} \le \|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)}\frac{M}{\alpha}\bigg(\sqrt{3}\big(M\|\eta_\tau\|_{\ell^2(J;V)} + \|\partial_t\eta\|_{L^2(J;V')} + \tau\|\partial_{tt}u\|_{L^2(J;V')}\big) + \sqrt{\alpha}\|e_h^0\|_L\bigg).$$

Furthermore, since $\eta(t_n) - \eta(t_{n-1}) = \int_{J_n}\partial_t\eta(s)\,\mathrm{d}s$, using the Cauchy–Schwarz inequality, and recalling that $(\delta_\tau u_\tau)^n := \frac{1}{\tau}(u(t_n) - u(t_{n-1}))$, we infer that

$$\begin{aligned}
\|\delta_\tau u_{h\tau} - \delta_\tau u_\tau\|_{\ell^2(J;V')}^2 &= \tau^{-2}\sum_{n\in\mathcal{N}_\tau}\tau\Big\|\int_{J_n}\partial_t\eta(s)\,\mathrm{d}s\Big\|_{V'}^2 \\
&\le \tau^{-2}\sum_{n\in\mathcal{N}_\tau}\tau\Big(\int_{J_n}\|\partial_t\eta(s)\|_{V'}\,\mathrm{d}s\Big)^2 \\
&\le \tau^{-2}\sum_{n\in\mathcal{N}_\tau}\tau^2\int_{J_n}\|\partial_t\eta(s)\|_{V'}^2\,\mathrm{d}s \\
&= \|\partial_t\eta\|_{L^2(J;V')}^2.
\end{aligned}$$

Using the triangle inequality for

$$\frac{(u(t_n)-u_h^n)-(u(t_{n-1})-u_h^{n-1})}{\tau}=\frac{\eta(t_n)-\eta(t_{n-1})}{\tau}-(\delta_\tau e_{h\tau})^n$$

leads to the assertion (notice that $\|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)}\frac{M}{\alpha}\geq 1$ and recall that $\|e_h^0\|_L\leq\|\eta(0\|_L)$.

Exercise 71.9 (Inf-sup for cPG(k)). Let $v_{h\tau}\in X_{h\tau}$ and recall that $y_{h\tau}\in Y_{h\tau}$ is s.t. $y_{h\tau}(0):=v_{h\tau}(0)$ and $y_{h\tau}(t):=\mathcal{I}_{k-1}^{\mathrm{GL}}(A_h^{-1}(\partial_t v_{h\tau})+v_{h\tau})(t)$ for all $t\in J_\tau$. Using the coercivity of A_h at $t_{n,l}$ for all $n\in\mathcal{N}_\tau$ and $l\in\{1{:}k{+}1\}$, we infer that $\alpha\|y_{h\tau}\|_{Y_{h\tau}}^2\leq\|v_{h\tau}(0)\|_L^2+\mathfrak{T}_3$, where

$$\begin{aligned}\mathfrak{T}_3&:=\int_J(A_h(y_{h\tau}),y_{h\tau})_L\,\mu_k^{\mathrm{GL}}(\mathrm{d}t)\\&=\int_J(A_h(A_h^{-1}(\partial_t v_{h\tau})+v_{h\tau}),A_h^{-1}(\partial_t v_{h\tau})+v_{h\tau})_L\,\mu_k^{\mathrm{GL}}(\mathrm{d}t),\end{aligned}$$

where we used (70.1b). Rearranging the terms and since $(\partial_t v_{h\tau},v_{h\tau})_L\in\mathbb{P}_{2k-1}(J_n;\mathbb{R})$ for all $n\in\mathcal{N}_\tau$, we obtain

$$\begin{aligned}\mathfrak{T}_3=&\int_J 2(\partial_t v_{h\tau},v_{h\tau})_L\mathrm{d}t+\int_J(A_h(v_{h\tau}),v_{h\tau})_L\,\mu_k^{\mathrm{GL}}(\mathrm{d}t)\\&+\int_J(A_h^{-1}(\partial_t v_{h\tau}),\partial_t v_{h\tau})_L\,\mu_k^{\mathrm{GL}}(\mathrm{d}t).\end{aligned}$$

Using the boundedness of $A_h(t_{n,l})$ with constant bounded by M, the boundedness of $A_h(t_{n,l})^{-1}$ with constant bounded by $\frac{1}{\alpha}$ for all $n\in\mathcal{N}_\tau$ and all $l\in\{1{:}k\}$, and observing that $2\int_J(\partial_t v_{h\tau},y_{h\tau})_L\mathrm{d}t=\|v_{h\tau}(T)\|_L^2-\|v_{h\tau}(0)\|_L^2$, we finally conclude that $\|y_{h\tau}\|_{Y_{h\tau}}^2\leq\frac{M}{\alpha}\|v_{h\tau}\|_{X_{h\tau}}^2$.

Chapter 72

Weak formulations and well-posedness

Exercises

Exercise 72.1 (Non-homogeneous Dirichlet condition). Consider the time-dependent Stokes equations (72.1) with the non-homogeneous Dirichlet condition $\boldsymbol{u} = \boldsymbol{g}$ enforced over the whole boundary ∂D for all $t \in J$. Assume that $\int_{\partial D} \boldsymbol{g}\cdot\boldsymbol{n} = 0$ for all $t \in J$. Assume that the data $\boldsymbol{f}$ and $\boldsymbol{g}$ are smooth so that the solution $(\boldsymbol{u}, p)$ is smooth. Assume that there is a smooth lifting $\boldsymbol{u}_{\boldsymbol{g}}$ of the boundary datum so that $\boldsymbol{u}_{\boldsymbol{g}}\cdot\boldsymbol{n} = \boldsymbol{g}$ on $\partial D \times J$ and $\nabla\cdot\boldsymbol{u}_{\boldsymbol{g}} = 0$ on $D \times J$. (i) Write the equations satisfied by $\boldsymbol{u}_0 := \boldsymbol{u} - \boldsymbol{u}_{\boldsymbol{g}}$. (ii) Verify that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}_0\|_{\boldsymbol{L}^2}^2 + 2\mu\|\mathbb{e}(\boldsymbol{u}_0)\|_{\mathbb{L}^2}^2 = (\boldsymbol{f}, \boldsymbol{u}_0)_{\boldsymbol{L}^2} - (\partial_t\boldsymbol{u}_{\boldsymbol{g}}, \boldsymbol{u}_0)_{\boldsymbol{L}^2} - 2\mu(\mathbb{e}(\boldsymbol{u}_{\boldsymbol{g}}), \mathbb{e}(\boldsymbol{u}_0))_{\mathbb{L}^2}.$$

(iii) Establish a priori bound on $\boldsymbol{u}_0$ of the form $\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}_0\|_{\boldsymbol{L}^2}^2 + 2\mu\|\mathbb{e}(\boldsymbol{u}_0)\|_{\mathbb{L}^2}^2 \le \Phi(T, \boldsymbol{f}, \boldsymbol{u}_{\boldsymbol{g}}) + \frac{1}{T}\|\boldsymbol{u}_0\|_{\boldsymbol{L}^2}^2$.

Exercise 72.2 (Space-time de Rham in L^2). (i) Show that the operator $\nabla\cdot : L^2(J; \boldsymbol{H}_0^1(D)) \to L^2(J; L^2_*(D))$ is surjective. (*Hint*: invoke Lemma 53.9, Lemma C.44, and Corollary 64.14.) (ii) Show that $\boldsymbol{S} \in L^2(J; \boldsymbol{H}^{-1}(D))$ satisfies $\int_J \langle \boldsymbol{S}, \boldsymbol{w}\rangle_{\boldsymbol{H}^{-1},\boldsymbol{H}_0^1}\mathrm{d}t = 0$ for all $\boldsymbol{w} \in L^2(J; \boldsymbol{\mathcal{V}})$ iff there is $p \in L^2(J; L^2_*(D))$ s.t. $\int_J \langle \boldsymbol{S}, \boldsymbol{w}\rangle_{\boldsymbol{H}^{-1},\boldsymbol{H}_0^1}\mathrm{d}t = \int_J (p, \nabla\cdot\boldsymbol{w})_{L^2}\mathrm{d}t$ for all $\boldsymbol{w} \in L^2(J; \boldsymbol{H}_0^1(D))$. (*Hint*: use the closed range theorem.)

Exercise 72.3 (Variable viscosity). Assume that μ depends on $\boldsymbol{x} \in D$, and set $0 < \mu_\flat := \operatorname{ess\,inf}_{\boldsymbol{x}\in D}\mu$, $\mu_\sharp := \operatorname{ess\,sup}_{\boldsymbol{x}\in D}\mu < \infty$. Consider the mixed weak formulation (72.12). Prove that $\mu_\flat\|\boldsymbol{u}\|_{L^2(J;\boldsymbol{\mathcal{V}})}^2 \le \frac{1}{4}\rho\|\boldsymbol{f}\|_{L^2(J;\boldsymbol{L}^2)}^2 + \frac{1}{2}\|\boldsymbol{u}_0\|_{\boldsymbol{L}^2}^2$ with $\rho := C_{\mathrm{KPS}}^{-2}\frac{\ell_D^2}{\mu_\flat}$, $\|\partial_t\boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}^2 \le \|\boldsymbol{f}\|_{L^2(J;\boldsymbol{L}^2)}^2 + 2\mu_\sharp\|\boldsymbol{u}_0\|_{\boldsymbol{\mathcal{V}}}^2$, and $\|p\|_{L^2(J;L^2)}^2 \le \frac{1}{\beta^2}\big(c_1\|\boldsymbol{f}\|_{L^2(J;\boldsymbol{L}^2)}^2 + c_2\|\boldsymbol{u}_0\|_{\boldsymbol{\mathcal{V}}}^2\big)$ with $c_1 := \rho\mu_\flat(8+2\xi_\mu^2)$, $c_2 := \rho\mu_\flat\mu_\sharp(8+4\xi_\mu)$, and $\xi_\mu := \frac{\mu_\sharp}{\mu_\flat}$. (*Hint*: adapt the proof of Theorem 72.3.)

Exercise 72.4 (Distributional time derivative). Let $V \hookrightarrow L \equiv L' \hookrightarrow V'$ be a Gelfand triple. (i) Let $v \in X(J; V, V')$. Show that the action of $\hat\partial_t v \in H^{-1}(J; V')$ and of $\partial_t v \in L^2(J; V')$ coincide on $H_0^1(J; V)$. (*Hint*: use the integration by parts formula from Lemma 64.40.) (ii) Let $v \in H^1(J; L)$. Show that the action of $\hat\partial_t v \in H^{-1}(J; V')$ and of $\partial_t v \in L^2(J; L)$ coincide on $H_0^1(J; V)$. (*Hint*: as above.)

Exercise 72.5 (Space-time de Rham in H^{-1}). (i) Show that the operator $\nabla\cdot : H^1(J;\boldsymbol{H}^1_0(D)) \to H^1(J;L^2_*(D))$ is surjective. (*Hint*: proceed as in Exercise 72.2 and use Lemma 64.34.) (ii) Show that $\nabla\cdot : H^1_0(J;\boldsymbol{H}^1_0(D)) \to H^1_0(J;L^2_*(D))$ is surjective. (*Hint*: use Step (i) and Lemma 64.37.) (iii) Prove Lemma 72.8. (*Hint*: use the closed range theorem.)

Solution to exercises

Exercise 72.1 (Non-homogeneous Dirichlet condition). (i) Using the decomposition $\boldsymbol{u} := \boldsymbol{u}_0 + \boldsymbol{u_g}$ and the properties of the lifting $\boldsymbol{u_g}$, we infer that the governing equations are

$$\begin{aligned}
&\partial_t \boldsymbol{u}_0 - 2\mu\nabla\cdot\mathbb{e}(\boldsymbol{u}_0) + \nabla p = \boldsymbol{f} - \partial_t \boldsymbol{u_g} + 2\mu\nabla\cdot\mathbb{e}(\boldsymbol{u_g}) && \text{in } D\times J,\\
&\nabla\cdot\boldsymbol{u} = 0 && \text{in } D\times J,\\
&\boldsymbol{u}_0 = \boldsymbol{0} && \text{on } \partial D\times J,\\
&\boldsymbol{u}_0(\cdot,0) = \boldsymbol{u}_0(\cdot) - \boldsymbol{u_g}(\cdot,0) && \text{in } D.
\end{aligned}$$

(ii) Multiplying the momentum equation by $\boldsymbol{u}_0$, integrating over D, using that $\nabla\cdot\boldsymbol{u}_0 = 0$ to cancel the term involving the pressure and that $\boldsymbol{u}_0 = \boldsymbol{0}$ on ∂D to integrate by parts the terms involving the divergence of the linearized strain tensor, we obtain the assertion:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2} + 2\mu\|\mathbb{e}(\boldsymbol{u}_0)\|^2_{\mathbb{L}^2} = (\boldsymbol{f},\boldsymbol{u}_0)_{\boldsymbol{L}^2} - (\partial_t\boldsymbol{u_g},\boldsymbol{u}_0)_{\boldsymbol{L}^2} - 2\mu(\mathbb{e}(\boldsymbol{u_g}),\mathbb{e}(\boldsymbol{u}_0))_{\mathbb{L}^2}.$$

(iii) Invoking the Cauchy–Schwarz inequality and Young's inequality leads to

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2} + 2\mu\|\mathbb{e}(\boldsymbol{u}_0)\|^2_{\mathbb{L}^2} \le \frac{T}{2}\|\boldsymbol{f} - \partial_t\boldsymbol{u_g}\|^2_{\boldsymbol{L}^2} + \frac{1}{2T}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2} + \mu\|\mathbb{e}(\boldsymbol{u_g})\|^2_{\mathbb{L}} + \mu\|\mathbb{e}(\boldsymbol{u}_0)\|^2_{\mathbb{L}}.$$

Rearranging the terms gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2} + 2\mu\|\mathbb{e}(\boldsymbol{u}_0)\|^2_{\mathbb{L}^2} \le T\|\boldsymbol{f} - \partial_t\boldsymbol{u_g}\|^2_{\boldsymbol{L}^2} + 2\mu\|\mathbb{e}(\boldsymbol{u_g})\|^2_{\mathbb{L}} + \frac{1}{T}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2}.$$

Notice that all the terms have consistent dimensions.

Exercise 72.2 (Space-time de Rham in L^2). (i) Owing to Lemma 53.9, the linear operator $\nabla\cdot : \boldsymbol{H}^1_0(D) \to L^2_*(D)$ is surjective. Then, according to Lemma C.44, this operator has a linear right inverse. Let us denote by $\mathrm{div}^\dagger$ the right inverse in question. Thus, there exists $c > 0$ such that $\nabla\cdot(\mathrm{div}^\dagger(r)) = r$ and $\|\mathrm{div}^\dagger(r)\|_{\boldsymbol{V}} \le c\|r\|_{L^2(D)}$ for all $r \in L^2_*(D)$. Let now $q \in L^2(J;L^2_*(D))$. Corollary 64.14 and the linearity of $\mathrm{div}^\dagger : L^2_*(D) \to \boldsymbol{H}^1_0(D)$ imply that $\mathrm{div}^\dagger(q)$ is Bochner integrable. Moreover, $\|\mathrm{div}^\dagger(q(t))\|_{\boldsymbol{V}} \le c\|q(t)\|_{L^2(D)}$ for a.e. $t \in J$. We infer that

$$\|\mathrm{div}^\dagger(q)\|_{L^2(J;\boldsymbol{V})} \le c\,\|q\|_{L^2(J;L^2)}.$$

In conclusion, $\nabla\cdot(\mathrm{div}^\dagger(q(t))) = q(t)$ for a.e. $t \in J$, and $\mathrm{div}^\dagger(q) \in L^2(J;\boldsymbol{H}^1_0(D))$. This proves that $\nabla\cdot : L^2(J;\boldsymbol{H}^1_0(D)) \to L^2(J;L^2_*(D))$ is surjective.
(ii) Since $\nabla\cdot : L^2(J;\boldsymbol{H}^1_0(D)) \to L^2(J;L^2_*(D))$ is surjective, its range is closed. Notice that $(\nabla\cdot)^* : L^2(J;L^2_*(D)) \to L^2(J;\boldsymbol{H}^{-1}(D))$ since $(L^2_*(D))' \equiv L^2_*(D)$ and $(L^2(J;\boldsymbol{H}^1_0(D)))' = L^2(J;\boldsymbol{H}^{-1}(D))$ owing to Lemma 64.20(i). The closed range theorem (Theorem C.35) implies that $(\ker(\nabla\cdot))^\perp = \mathrm{im}((\nabla\cdot)^*)$ (here, we use the annihilator notation introduced in (C.14a)). Since $\ker(\nabla\cdot) = L^2(J;\boldsymbol{\mathcal{V}})$

(indeed, $\boldsymbol{v} \in L^2(J;\boldsymbol{\mathcal{V}})$ iff $\boldsymbol{v} \in L^2(J;\boldsymbol{H}_0^1(D))$ and $\nabla\cdot\boldsymbol{v} = 0$ in $L^2(J)$) and since our assumption on the linear form $\boldsymbol{S} \in L^2(J;\boldsymbol{H}^{-1}(D))$ means that $\boldsymbol{S} \in (\ker(\nabla\cdot))^\perp$, we infer that there is $p \in L^2(J;L^2_*(D))$ s.t. $\boldsymbol{S} = (\nabla\cdot)^*(p)$. This means that we have for all $\boldsymbol{w} \in L^2(J;\boldsymbol{H}_0^1(D))$,

$$\begin{aligned}\langle \boldsymbol{S}, \boldsymbol{w}\rangle_{L^2(\boldsymbol{H}^{-1}),L^2(\boldsymbol{H}_0^1)} &= \langle (\nabla\cdot)^*(p), \boldsymbol{w}\rangle_{L^2(\boldsymbol{H}^{-1}),L^2(\boldsymbol{H}_0^1)}\\ &= (p, \nabla\cdot\boldsymbol{w})_{L^2(J;L^2_*(D))} = \int_J (p(t), \nabla\cdot\boldsymbol{w}(t))_{L^2}\mathrm{d}t.\end{aligned}$$

Exercise 72.3 (Variable viscosity). Using the same arguments as in Step (1) of the proof of Theorem 72.3, we infer that

$$\langle \partial_t\boldsymbol{u}(t), \boldsymbol{u}(t)\rangle_{\boldsymbol{\mathcal{V}}',\boldsymbol{\mathcal{V}}} + \mu_\flat\|\boldsymbol{u}(t)\|^2_{\boldsymbol{\mathcal{V}}} \le \frac{C_{\mathrm{KPS}}^2}{4\mu_\flat\ell_D^2}\|\boldsymbol{f}(t)\|^2_{\boldsymbol{L}^2}.$$

Using again the same arguments leads to the estimate on $\|\boldsymbol{u}\|_{L^2(J;\boldsymbol{\mathcal{V}})}$. The bound on $\|\partial_t\boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}$ is derived by repeating the arguments from Step (2) of this proof, the only difference being that invoking the boundedness of a now yields the estimate

$$\|\partial_t\boldsymbol{u}_n\|^2_{L^2(J;\boldsymbol{L}^2)} \le \|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + 2\mu_\sharp\|\boldsymbol{u}_0\|^2_{\boldsymbol{\mathcal{V}}}.$$

Finally, to estimate the pressure, we can still proceed as in Step (3) and use the above estimates on $\|\partial_t\boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}$ and on $\|\boldsymbol{u}\|_{L^2(J;\boldsymbol{\mathcal{V}})}$. This yields

$$\beta\|p(t)\|_{L^2(D)} \le \frac{\ell_D}{C_{\mathrm{KPS}}}\big(\|\partial_t\boldsymbol{u}(t)\|_{\boldsymbol{L}^2(D)} + \|\boldsymbol{f}(t)\|_{\boldsymbol{L}^2(D)}\big) + 2\mu_\sharp\|\boldsymbol{u}(t)\|_{\boldsymbol{\mathcal{V}}}.$$

Squaring, using the definition of the time scale ρ, and integrating over time leads to

$$\beta^2\|p\|^2_{L^2(J;L^2)} \le 4\rho\mu_\flat\big(2\|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + 2\mu_\sharp\|\boldsymbol{u}_0\|^2_{\boldsymbol{\mathcal{V}}}\big) + 8\frac{\mu_\sharp^2}{\mu_\flat}\big(\tfrac14\rho\|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + \tfrac12\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2}\big),$$

which leads to the expected bound on the pressure after observing that $\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2} \le \frac{\ell_D^2}{C_{\mathrm{KPS}}^2}\|\boldsymbol{u}_0\|^2_{\boldsymbol{\mathcal{V}}} = \mu_\flat\rho\|\boldsymbol{u}_0\|^2_{\boldsymbol{\mathcal{V}}}$ and rearranging the terms.

Exercise 72.4 (Distributional time derivative). (i) Let $v \in X(J;V,V')$. Since $V \hookrightarrow V'$, we have $v \in L^2(J;V')$ so that it is meaningful to define the distributional time derivative $\hat\partial_t v \in H^{-1}(J;V')$, and we have

$$\langle \hat\partial_t v, w\rangle_{H^{-1}(V'),H_0^1(V)} := -\int_J \langle v, \partial_t w\rangle_{V',V}\mathrm{d}t,$$

for all $w \in H_0^1(J;V)$. Moreover, since both v and w are in $X(J;V,V')$ and since $w(0) = w(T) = 0$ by assumption, the integration by parts formula from Lemma 64.40 implies that

$$\int_J \langle \partial_t v, w\rangle_{V',V}\mathrm{d}t = -\int_J \langle v, \partial_t w\rangle_{V,V'}\mathrm{d}t.$$

This shows that

$$\langle \hat\partial_t v, w\rangle_{H^{-1}(V'),H_0^1(V)} = \int_J \langle \partial_t v, w\rangle_{V',V}\mathrm{d}t,$$

for all $w \in H_0^1(J;V)$. Notice in passing that since the duality product between V' and V is an extension of the L-inner product, we have

$$\langle \hat\partial_t v, w\rangle_{H^{-1}(V'),H_0^1(V)} = \int_J \langle \partial_t v, w\rangle_{V',V}\mathrm{d}t = -\int_J (v, \partial_t w)_L\mathrm{d}t.$$

(ii) A short answer consists of saying that the assertion follows from Step (i) since $H^1(J;L) \hookrightarrow X(J;L;L)$. (Here, the Gelfand triple is simply $L \hookrightarrow L \equiv L' \hookrightarrow L'$.) One can also answer the question by redoing the proof. Let $v \in H^1(J;L)$. Then $v \in L^2(J;L) \hookrightarrow L^2(J;V')$, so that it is meaningful to define the distributional time derivative $\hat{\partial}_t v$, and we have by definition

$$\langle \hat{\partial}_t v, w\rangle_{H^{-1}(V'),H_0^1(V)} := -\int_J \langle v, \partial_t w\rangle_{V',V} \mathrm{d}t = -\int_J (v, \partial_t w)_L \mathrm{d}t,$$

for all $w \in H_0^1(J;V)$ since $v \in L^2(J;L)$. Moreover, since the functions v and w are in $H^1(J;L) = X(J;L,L)$ and $w(0) = w(T) = 0$, the integration by parts formula from Lemma 64.40 implies that $-\int_J (v, \partial_t w)_L \mathrm{d}t = \int_J (\partial_t v, w)_L \mathrm{d}t = (\partial_t v, w)_{L^2(J;L)}$. In conclusion, we have shown that

$$\langle \hat{\partial}_t v, w\rangle_{H^{-1}(V'),H_0^1(V)} = (\partial_t v, w)_{L^2(J;L)}, \qquad \forall w \in H_0^1(J;V).$$

Exercise 72.5 (Space-time de Rham in H^{-1}). (i) We consider the right inverse operator $\mathrm{div}^\dagger : L^2_*(D) \to \boldsymbol{H}^1(D)$ introduced in Exercise 72.2. Let $q \in H^1(J;L^2_*(D))$. We have already shown that $\mathrm{div}^\dagger(q) \in L^2(J;\boldsymbol{H}_0^1(D))$ where $\mathrm{div}^\dagger(q)(t) = \mathrm{div}^\dagger(q(t))$ for a.e. $t \in J$. Moreover, since $\partial_t q \in L^2(J;L^2_*(D))$ by assumption, we also have $\mathrm{div}^\dagger(\partial_t q) \in L^2(J;\boldsymbol{H}_0^1(D))$ with

$$\|\mathrm{div}^\dagger(\partial_t q)\|_{L^2(J;\boldsymbol{V})} \le c\,\|\partial_t q\|_{L^2(J;L^2)}.$$

Finally, using the linearity of $\mathrm{div}^\dagger$ and Lemma 64.34, we infer that $\mathrm{div}^\dagger(\partial_t q) = \partial_t \mathrm{div}^\dagger(q)$. Hence, we have

$$\|\partial_t \mathrm{div}^\dagger(q)\|_{L^2(J;\boldsymbol{V})} = \|\mathrm{div}^\dagger(\partial_t q)\|_{L^2(J;\boldsymbol{V})} \le c\,\|\partial_t q\|_{L^2(J;L^2)}.$$

In conclusion, $\nabla\cdot(\mathrm{div}^\dagger(q(t))) = q(t)$ for a.e. $t \in J$, and $\|\mathrm{div}^\dagger(q)\|_{H^1(J;\boldsymbol{H}_0^1(D))} \le c\|q\|_{H^1(J;L^2_*(D))}$ for all $q \in H^1(J;L^2_*(D))$. Hence, $\mathrm{div}^\dagger : H^1(J;L^2_*(D)) \to H^1(J;\boldsymbol{H}_0^1(D))$ is a right inverse of $\nabla\cdot : H^1(J;\boldsymbol{H}_0^1(D)) \to H^1(J;L^2_*(D))$.
(ii) From the above argument, we deduce that for all $q \in H_0^1(J;L^2_*(D))$, there exists $\boldsymbol{v} := \mathrm{div}^\dagger(q) \in H^1(J;L^2_*(D))$. But $\|\boldsymbol{v}(t)\|_{\boldsymbol{V}} \le c\|q(t)\|_{L^2(D)}$, and this inequality holds true for every $t \in \overline{J}$ since $q \in C^{0,\frac12}(\overline{J};L^2_*(D))$ owing to Lemma 64.37. This implies that $\boldsymbol{v}(0) = \boldsymbol{v}(T) = \boldsymbol{0}$. We infer that $\boldsymbol{v} \in H_0^1(J;\boldsymbol{H}_0^1(D))$. We have proved that $\mathrm{div}^\dagger : H_0^1(J;L^2_*(D)) \to H_0^1(J;\boldsymbol{H}_0^1(D))$ is a right inverse of $\nabla\cdot : H_0^1(J;\boldsymbol{H}_0^1(D)) \to H_0^1(J;L^2_*(D))$.
(iii) From Step (ii), we know that the operator $\nabla\cdot : H_0^1(J;\boldsymbol{H}_0^1(D)) \to H_0^1(J;L^2_*(D))$ is surjective so that its range is closed. Notice that $(\nabla\cdot)^* : H^{-1}(J;L^2_*(D)) \to H^{-1}(J;\boldsymbol{H}^{-1}(D))$. The closed range theorem (Theorem C.35) implies that $(\ker(\nabla\cdot))^\perp = \mathrm{im}((\nabla\cdot)^*)$ (here, we use the annihilator notation introduced in (C.14a)). Furthermore, $\boldsymbol{v} \in H_0^1(J;\boldsymbol{\mathcal{V}})$ iff $\boldsymbol{v} \in H_0^1(J;\boldsymbol{H}_0^1(D))$ and $\nabla\cdot\boldsymbol{v} = 0$ in $L^2(J)$. Hence, $H_0^1(J;\boldsymbol{\mathcal{V}}) = \ker(\nabla\cdot)$ with $\nabla\cdot : H_0^1(J;\boldsymbol{H}_0^1(D)) \to H_0^1(J;L^2_*(D))$. Since our assumption on the linear form $\boldsymbol{S} \in H^{-1}(J;\boldsymbol{H}^{-1}(D))$ means that $\boldsymbol{S} \in (\ker(\nabla\cdot))^\perp$, and invoking the identity $(\ker(\nabla\cdot))^\perp = \mathrm{im}((\nabla\cdot)^*)$, we infer that there is $p \in H^{-1}(J;L^2_*(D))$ s.t. $\boldsymbol{S} = (\nabla\cdot)^*(p)$. This means that we have for all $\boldsymbol{w} \in H_0^1(J;\boldsymbol{H}_0^1(D))$,

$$\begin{aligned}\langle \boldsymbol{S}, \boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H_0^1(\boldsymbol{H}_0^1)} &= \langle (\nabla\cdot)^*(p), \boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H_0^1(\boldsymbol{H}_0^1)} \\ &= \langle p, \nabla\cdot\boldsymbol{w}\rangle_{H^{-1}(L^2);H_0^1(L^2)}.\end{aligned}$$

Chapter 73

Monolithic time discretization

Exercises

Exercise 73.1 (Well-posedness). Prove Proposition 73.1. (*Hint*: adapt the proof of Theorem 72.3.)

Exercise 73.2 (Simplified Gronwall's lemma). Let $a \in W^{1,1}(J;\mathbb{R})$, let $b \in L^\infty(\overline{J};\mathbb{R})$, and let $\gamma > 0$. Assume that $\frac{\mathrm{d}}{\mathrm{d}t}a(t) \le \frac{1}{\gamma}a(t) + b(t)$ for all $t \in \overline{J}$. Prove that $a(t) \le e^{\frac{t}{\gamma}}\big(a(0) + \min(t,\gamma)\|b\|_{L^\infty(\overline{J}_t)}\big)$ with $\overline{J}_t := (0,t)$ for all $t \in \overline{J}$. (*Hint*: observe that $\int_0^t e^{\frac{t-s}{\gamma}}\,\mathrm{d}s \le \min(t,\gamma)e^{\frac{t}{\gamma}}$.) *Note*: this is a simplified form of Gronwall's lemma; see Exercise 65.3.

Exercise 73.3 (BDF2, Crank–Nicolson). (i) Using the setting described in §68.2 for BDF2, write the discrete formulation and the algebraic realization of the time-dependent Stokes equations with the time discretization performed with BDF2. (ii) Same question for the Crank–Nicolson scheme using the setting described in §68.3. (iii) Same question for the Crank–Nicolson scheme using the setting described in §73.4.

Solution to exercises

Exercise 73.1 (Well-posedness). One first proves by means of the Cauchy–Lipschitz theorem that there is a unique $\boldsymbol{u}_h \in H^1(J;\boldsymbol{\mathcal{V}}_h)$ such that

$$(\partial_t \boldsymbol{u}_h(t), \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{u}_h(t), \boldsymbol{w}_h) = (\boldsymbol{f}(t), \boldsymbol{w}_h)_{\boldsymbol{L}^2},$$

in $L^2(J)$ for all $\boldsymbol{w}_h \in \boldsymbol{\mathcal{V}}_h$. Then one infers the existence and uniqueness of the pressure in $L^2(J;Q_h)$ by invoking the inf-sup condition (73.4) and reasoning as in Exercise 72.2.

Exercise 73.2 (Simplified Gronwall's lemma). Rearranging the terms, we infer that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(e^{-\frac{t}{\gamma}}a(t)\Big) \le e^{-\frac{t}{\gamma}}b(t),$$

for all $t \in \overline{J}$. Integrating from 0 to t gives

$$a(t) \le e^{\frac{t}{\gamma}} a(0) + \int_0^t e^{\frac{t-s}{\gamma}} b(s)\,\mathrm{d}s \le e^{\frac{t}{\gamma}} a(0) + \left(\int_0^t e^{\frac{t-s}{\gamma}}\,\mathrm{d}s \right) \|b\|_{L^\infty(J_t)}.$$

The assertion follows by observing that $\int_0^t e^{\frac{t-s}{\gamma}}\,\mathrm{d}s \le t e^{\frac{t}{\gamma}}$ on the one hand and that $\int_0^t e^{\frac{t-s}{\gamma}}\,\mathrm{d}s = \gamma(e^{\frac{t}{\gamma}} - 1) \le \gamma e^{\frac{t}{\gamma}}$ on the other hand.

Exercise 73.3 (BDF2, Crank–Nicolson). (i) For the BDF2 scheme, we first set $\boldsymbol{u}_h^0 := \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{u}_0, 0)$, then we compute $(\boldsymbol{u}_h^n, p_h^n) \in \boldsymbol{V}_h \times Q_h$ for all $n \in \mathcal{N}_\tau$ so that the following holds true: For $n = 1$,

$$\begin{cases} \dfrac{1}{\tau}(\boldsymbol{u}_h^1 - \boldsymbol{u}_h^0, \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{u}_h^1, \boldsymbol{w}_h) + b(\boldsymbol{w}_h, p_h^1) = (\boldsymbol{f}^1, \boldsymbol{w}_h)_{\boldsymbol{L}^2}, \\ b(\boldsymbol{u}_h^1, q_h) = 0, \end{cases}$$

for all $(\boldsymbol{w}_h, q_h) \in \boldsymbol{V}_h \times Q_h$, and for all $n \ge 2$,

$$\begin{cases} \dfrac{1}{2\tau}(3\boldsymbol{u}_h^n - 4\boldsymbol{u}_h^{n-1} + 3\boldsymbol{u}_h^{n-2}, \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{u}_h^n, \boldsymbol{w}_h) + b(\boldsymbol{w}_h, p_h^n) = (\boldsymbol{f}^n, \boldsymbol{w}_h)_{\boldsymbol{L}^2}, \\ b(\boldsymbol{u}_h^n, q_h) = 0, \end{cases}$$

for all $(\boldsymbol{w}_h, q_h) \in \boldsymbol{V}_h \times Q_h$. The algebraic realization of the first time step can be written as follows:

$$\begin{pmatrix} \mathcal{M} + \tau\mathcal{A} & \mathcal{B}^{\mathsf{T}} \\ \mathcal{B} & \mathbb{O}_{K,K} \end{pmatrix} \begin{pmatrix} \mathsf{U}^1 \\ \mathsf{P}^1 \end{pmatrix} = \begin{pmatrix} \tau\mathsf{F}^1 + \mathcal{M}\mathsf{U}^0 \\ 0 \end{pmatrix}, \tag{73.1}$$

where $\mathsf{F}^1 := ((\boldsymbol{f}^1, \varphi_i)_{\boldsymbol{L}^2})_{i \in \{1:I\}}$ and $\mathbb{O}_{K,K}$ is the zero matrix in $\mathbb{R}^{K \times K}$. The other time steps give

$$\begin{pmatrix} \frac{3}{2}\mathcal{M} + \tau\mathcal{A} & \mathcal{B}^{\mathsf{T}} \\ \mathcal{B} & \mathbb{O}_{K,K} \end{pmatrix} \begin{pmatrix} \mathsf{U}^n \\ \mathsf{P}^n \end{pmatrix} = \begin{pmatrix} \tau\mathsf{F}^n + \mathcal{M}(2\mathsf{U}^{n-1} - \frac{1}{2}\mathsf{U}^{n-2}) \\ 0 \end{pmatrix},$$

where $\mathsf{F}^n := ((\boldsymbol{f}^n, \varphi_i)_{\boldsymbol{L}^2})_{i \in \{1:I\}}$.

(ii) For the Crank–Nicolson scheme, we first set $\boldsymbol{u}_h^0 := \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{u}_0, 0)$, then we compute $(\boldsymbol{u}_h^n, p_h^{n-\frac{1}{2}}) \in \boldsymbol{V}_h \times Q_h$ for all $n \in \mathcal{N}_\tau$ so that the following holds true:

$$\begin{cases} \dfrac{1}{\tau}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\frac{1}{2}(\boldsymbol{u}_h^n + \boldsymbol{u}_h^{n-1}), \boldsymbol{w}_h) + b(\boldsymbol{w}_h, p_h^{n-\frac{1}{2}}) = (\boldsymbol{f}^{n-\frac{1}{2}}, \boldsymbol{w}_h)_{\boldsymbol{L}^2}, \\ b(\boldsymbol{u}_h^n, q_h) = 0, \end{cases}$$

for all $(\boldsymbol{w}_h, q_h) \in \boldsymbol{V}_h \times Q_h$. The algebraic realization of the scheme can be written as follows:

$$\begin{pmatrix} \mathcal{M} + \frac{\tau}{2}\mathcal{A} & \mathcal{B}^{\mathsf{T}} \\ \mathcal{B} & \mathbb{O}_{K,K} \end{pmatrix} \begin{pmatrix} \mathsf{U}^n \\ \mathsf{P}^{n-\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} \tau\mathsf{F}^{n-\frac{1}{2}} + \mathcal{M}\mathsf{U}^{n-1} - \frac{\tau}{2}\mathcal{A}\mathsf{U}^{n-1} \\ 0 \end{pmatrix},$$

where $\mathsf{F}^{n-\frac{1}{2}}$ is the coordinate vector of $\boldsymbol{f}^{n-\frac{1}{2}}$, $\mathsf{P}^{n-\frac{1}{2}}$ is the coordinate vector of $\tau p_h^{n-\frac{1}{2}}$, and $p_h^{n-\frac{1}{2}}$ is the approximation of $p(t_n - \frac{\tau}{2}) = p(\frac{t_{n-1}+t_n}{2})$.

(iii) Adopting the point of view from §73.4, we obtain the linear system

$$\begin{pmatrix} \mathcal{M} + \frac{\tau}{2}\mathcal{A} & \mathcal{B}^{\mathsf{T}} \\ \mathcal{B} & \mathbb{O}_{K,K} \end{pmatrix} \begin{pmatrix} \mathsf{U}^{n,1} \\ \mathsf{P}^{n,1} \end{pmatrix} = \begin{pmatrix} \mathcal{M}\mathsf{U}^{n-1} + \frac{1}{2}\tau\mathsf{F}^{n-\frac{1}{2}} \\ 0 \end{pmatrix},$$

and $\mathsf{U}^n = 2\mathsf{U}^{n,1} - \mathsf{U}^{n-1}$, i.e., $\mathsf{U}^{n,1} = \frac{1}{2}(\mathsf{U}^n + \mathsf{U}^{n-1})$. This is equivalent to the linear system obtained in Step (ii), once we set $\mathsf{P}^{n-\frac{1}{2}} := 2\mathsf{P}^{n,1}$, which is consistent since $\mathsf{P}^{n-\frac{1}{2}}$ approximates $\tau p(t_n - \frac{\tau}{2})$ and $\mathsf{P}^{n,1}$ approximates $\frac{1}{2}\tau p(t_{n,1}) = \frac{1}{2}\tau p(t_n - \frac{\tau}{2})$ (recall that $c_1 = \frac{1}{2}$).

Chapter 74

Projection methods

Exercises

Exercise 74.1 (Remark 74.1). Prove the stability estimate in Remark 74.6. (*Hint*: adapt the proof of Lemma 74.5.)

Exercise 74.2 (Curl-div-grad identity). Let $d \in \{2,3\}$. Show that $\|\nabla\times\boldsymbol{v}\|^2_{\boldsymbol{L}^2(D)}+\|\nabla\cdot\boldsymbol{v}\|^2_{L^2(D)} = \|\nabla\boldsymbol{v}\|^2_{\mathbb{L}^2(D)}$ for all $\boldsymbol{v}\in\boldsymbol{H}^1_0(D)$. (*Hint:* use $-\Delta\boldsymbol{v} = -\nabla(\nabla\cdot\boldsymbol{v}) + \nabla\times(\nabla\times\boldsymbol{v})$.)

Exercise 74.3 (Inverse of the Stokes operator). Let $\boldsymbol{V} := \boldsymbol{H}^1_0(D)$, $\boldsymbol{V}' = \boldsymbol{H}^{-1}(D)$, and $Q := L^2_*(D)$. The inverse of the Stokes operator $\boldsymbol{S} : \boldsymbol{H}^{-1}(D) \longrightarrow \boldsymbol{\mathcal{V}} := \{\boldsymbol{v}\in\boldsymbol{H}^1_0(D) \mid \nabla\cdot\boldsymbol{v} = 0\}$ is s.t. for all $\boldsymbol{f}\in\boldsymbol{V}'$, $\boldsymbol{S}(\boldsymbol{f})$ is the unique member of $\boldsymbol{\mathcal{V}}$ s.t. the following holds true for all $(\boldsymbol{w},q)\in\boldsymbol{V}\times Q$:

$$\begin{cases} 2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{f})),\mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2(D)} - (r,\nabla\cdot\boldsymbol{w})_{L^2(D)} = \langle\boldsymbol{f},\boldsymbol{w}\rangle_{\boldsymbol{V}',\boldsymbol{V}},\\ (q,\nabla\cdot\boldsymbol{S}(\boldsymbol{f}))_{L^2(D)} = 0,\end{cases}$$

where $\langle\cdot,\cdot\rangle_{\boldsymbol{V}',\boldsymbol{V}}$ denotes the duality pairing between $\boldsymbol{V}'$ and $\boldsymbol{V}$. Recall that $\mu\|\boldsymbol{S}(\boldsymbol{f})\|_{\boldsymbol{V}} + \|r\|_{L^2} \le c\|\boldsymbol{f}\|_{\boldsymbol{H}^{-1}}$ for all $\boldsymbol{f}\in\boldsymbol{H}^{-1}(D)$ with $\|\boldsymbol{w}\|_{\boldsymbol{V}} := \|\mathbb{e}(\boldsymbol{w})\|_{\mathbb{L}^2(D)}$. We assume that D is such that the following regularity property holds true: $\mu|\boldsymbol{S}(\boldsymbol{f})|_{\boldsymbol{H}^2} + |r|_{H^1} \le c\|\boldsymbol{f}\|_{\boldsymbol{L}^2}$ for all $\boldsymbol{f}\in\boldsymbol{L}^2(D)$. (i) Show that $2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{v})),\mathbb{e}(\boldsymbol{v}))_{\mathbb{L}^2} = \|\boldsymbol{v}\|^2_{\boldsymbol{L}^2}$ for all $\boldsymbol{v}\in\boldsymbol{\mathcal{V}}$. (*Hint*: recall that the duality pairing $\langle\cdot,\cdot\rangle_{\boldsymbol{V}',\boldsymbol{V}}$ is an extension of the $\boldsymbol{L}^2$-inner product.) (ii) Show that for all $\gamma\in(0,1)$, there is $c(\gamma)$ such that for all $\boldsymbol{v}$ in $\boldsymbol{V}$, $2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{v})),\mathbb{e}(\boldsymbol{v}))_{\mathbb{L}^2} \ge (1-\gamma)\|\boldsymbol{v}\|^2_{\boldsymbol{L}^2} - c(\gamma)\|\boldsymbol{v}-\boldsymbol{v}^\star\|^2_{\boldsymbol{L}^2}$ for all $\boldsymbol{v}^\star\in\boldsymbol{\mathcal{H}}$. (*Hint*: integrate by parts the pressure term.) (iii) Show that the map $\boldsymbol{V}'\ni\boldsymbol{v}\mapsto|\boldsymbol{v}|_\star := \langle\boldsymbol{v},\boldsymbol{S}(\boldsymbol{v})\rangle^{\frac12}_{\boldsymbol{V}',\boldsymbol{V}}$ defines a seminorm on $\boldsymbol{V}'$. Prove that $|\boldsymbol{v}|_\star \le (2\mu)^{-\frac12}\|\boldsymbol{v}\|_{\boldsymbol{V}'}$ for all $\boldsymbol{v}\in\boldsymbol{V}'$. *Note*: there does not exist any constant c so that $(2\mu)^{-\frac12}\|\boldsymbol{v}\|_{\boldsymbol{V}'} \le c|\boldsymbol{v}|_\star$ for all $\boldsymbol{v}\in\boldsymbol{H}^{-1}(D)$, i.e., $|\cdot|_\star$ is not a norm on $\boldsymbol{H}^{-1}(D)$; see Guermond [21, Thm. 4.1] and Guermond and Salgado [22, Thm. 32]. The inverse of the Stokes operator is used in Exercise 74.4 to prove Lemma 74.11.

Exercise 74.4 (Lemma 74.11). Consider the perturbed system (74.14), and set $\boldsymbol{e} := \boldsymbol{u}^\varepsilon - \boldsymbol{u}$ and $q := p^\varepsilon - p$. (i) Write the PDE system solved by the pair $(\boldsymbol{e},q)$ and show that

$$\frac12\frac{\mathrm{d}}{\mathrm{d}t}\|\partial_t\boldsymbol{e}\|^2_{\boldsymbol{L}^2} + 2\mu\|\partial_t\boldsymbol{e}\|^2_{\boldsymbol{V}} + \frac12\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla\phi^\varepsilon\|^2_{\boldsymbol{L}^2} + \frac12\varepsilon\mu\frac{\mathrm{d}}{\mathrm{d}t}\|\Delta\phi^\varepsilon\|^2_{L^2}$$
$$= \varepsilon\frac{\mathrm{d}}{\mathrm{d}t}(\nabla\partial_t p,\nabla\phi^\varepsilon)_{\boldsymbol{L}^2} - \varepsilon(\nabla\partial_{tt}p,\nabla\phi^\varepsilon)_{\boldsymbol{L}^2},$$

where we recall that $\boldsymbol{V} := \boldsymbol{H}^1_0(D)$ and $\|\boldsymbol{v}\|_{\boldsymbol{V}} := \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2}$. (ii) Prove that $\|\nabla\phi^\varepsilon(t)\|^2_{\boldsymbol{L}^2} \le c(p,T)\varepsilon^2$ for all $t \in J$. (*Hint*: use Gronwall's lemma from Exercise 65.3.) Conclude that $\|\nabla\cdot\boldsymbol{u}^\varepsilon\|^2_{L^\infty(J;L^2(D))} \le c(p,T)\mu^{-1}\varepsilon^3$. (iii) Show that $\|\boldsymbol{e} - \boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{e})\|^2_{\boldsymbol{L}^2} = \varepsilon^2\|\nabla\phi^\varepsilon\|^2_{\boldsymbol{L}^2}$, where the Leray projection $\boldsymbol{P}_{\boldsymbol{\mathcal{H}}}$ is defined in Lemma 74.1. Deduce from the above estimates that $\|\boldsymbol{u} - \boldsymbol{u}^\varepsilon\|_{L^2(J;\boldsymbol{L}^2(D))} \le c(p,T)\varepsilon^2$. (*Hint*: use the lower bound from Step (ii) of Exercise 74.3.)

Exercise 74.5 (Gauge-Uzawa). (i) Write the pressure-correction algorithm in rotational form using BDF1, $p^{\star,n} := p^{n-1}$, and the sequences $\widetilde{\boldsymbol{u}}_\tau \in (\boldsymbol{V})^N$, $\boldsymbol{u}_\tau \in (\boldsymbol{\mathcal{H}})^N$, $\phi_\tau \in (Q)^N$, $p_\tau \in (Q)^N$. (ii) Consider the sequences $\widetilde{\boldsymbol{v}}_\tau \in (\boldsymbol{V})^N$, $\boldsymbol{v}_\tau \in (\boldsymbol{V})^N$, $r_\tau \in (Q)^N$, $q_\tau \in (Q)^N$, $\psi_\tau \in (Q)^N$, generated by the following algorithm (called gauge-Uzawa in the literature, see Nochetto and Pyo [36]): Set $\boldsymbol{v}^0 := \boldsymbol{v}_0$, $r^0 := 0$, $q^0 = \psi^0 := p(0)$, then solve for all $n \in \mathcal{N}_\tau$,

$$
\begin{aligned}
&\frac{\widetilde{\boldsymbol{v}}^n - \boldsymbol{v}^{n-1}}{\tau} - \mu\Delta\widetilde{\boldsymbol{v}}^n + \nabla q^{n-1} = \boldsymbol{f}^n, \quad \widetilde{\boldsymbol{v}}^n_{|\partial D} = \boldsymbol{0},\\
&\boldsymbol{v}^n + \tau\nabla\psi^n = \widetilde{\boldsymbol{v}}^n + \tau\nabla\psi^{n-1}, \quad \nabla\cdot\boldsymbol{v}^n = 0, \quad \boldsymbol{v}^n_{|\partial D}\cdot\boldsymbol{n} = 0,\\
&r^n = r^{n-1} - \nabla\cdot\widetilde{\boldsymbol{v}}^n, \qquad\qquad q^n = \psi^n + \mu r^n.
\end{aligned}
$$

Recalling that $(\delta_\tau\psi_\tau)^n := \frac{\psi^n - \psi^{n-1}}{\tau}$ for all $n \in \mathcal{N}_\tau$, show that the sequences $(\widetilde{\boldsymbol{v}}_\tau, \boldsymbol{v}_\tau, \tau\delta_\tau\psi_\tau, q_\tau)$ and $(\widetilde{\boldsymbol{u}}_\tau, \boldsymbol{u}_\tau, \phi_\tau, p_\tau)$ are equal (i.e., the gauge-Uzawa and the pressure-correction method in rotational form are identical). (*Hint*: write $q^n = q^{n-1} + \psi^n - \psi^{n-1} + \mu(r^n - r^{n-1})$.) (iii) Show that for all $n \in \mathcal{N}_\tau$,

$$
\begin{aligned}
\|\boldsymbol{v}^n\|^2_{\boldsymbol{L}^2} + \tau^2\|\nabla\psi^n\|^2_{\boldsymbol{L}^2} + \mu\tau\|r^n\|^2_{L^2} + \|\widetilde{\boldsymbol{v}}^n - \boldsymbol{v}^{n-1}\|^2_{\boldsymbol{L}^2} + \frac{1}{2}\mu\tau\|\nabla\widetilde{\boldsymbol{v}}^n\|^2_{\mathbb{L}^2}&\\
\le \|\boldsymbol{v}^{n-1}\|^2_{\boldsymbol{L}^2} + \tau^2\|\nabla\psi^{n-1}\|^2_{\boldsymbol{L}^2} + \mu\tau\|r^{n-1}\|^2_{L^2} + \rho\tau\|\boldsymbol{f}^n\|^2_{\boldsymbol{L}^2},&
\end{aligned}
$$

with the time scale $\rho := \frac{2}{C^2_{\text{PS}}}\frac{\ell_D^2}{\mu}$. (*Hint*: test the momentum equation with $2\tau\widetilde{\boldsymbol{v}}^n$, square the second equation, square the third equation and scale the result by $\mu\tau$, and add the results.)

Solution to exercises

Exercise 74.1 (Remark 74.1). Testing (74.2) with $2\tau\widetilde{\boldsymbol{u}}^n$, using the coercivity of the bilinear form $a(\boldsymbol{v},\boldsymbol{w}) := (\mathbb{s}(\boldsymbol{v}), \mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2(D)}$ on $\boldsymbol{V}$, and the algebraic identity (67.9), we obtain

$$
\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{L}^2} - \|\boldsymbol{u}^{n-1}\|^2_{\boldsymbol{L}^2} + 4\mu\tau\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{V}} \le 2\tau(\boldsymbol{f}^n, \widetilde{\boldsymbol{u}}^n)_{\boldsymbol{L}^2}.
$$

Since

$$
2\tau(\boldsymbol{f}^n, \widetilde{\boldsymbol{u}}^n)_{\boldsymbol{L}^2} \le \frac{\tau}{2\mu}\|\boldsymbol{f}^n\|^2_{\boldsymbol{V}'} + 2\mu\tau\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{V}} \le \frac{\tau\rho}{2}\|\boldsymbol{f}^n\|^2_{\boldsymbol{L}^2} + 2\mu\tau\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{V}},
$$

where we used Young's inequality, the bound $\|\boldsymbol{f}^n\|_{\boldsymbol{V}'} \le C^{-1}_{\text{KPS}}\ell_D\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}$, and the definition of the time scale ρ, we infer that

$$
\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{L}^2} - \|\boldsymbol{u}^{n-1}\|^2_{\boldsymbol{L}^2} + 2\mu\tau\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{V}} \le \frac{\tau\rho}{2}\|\boldsymbol{f}^n\|^2_{\boldsymbol{L}^2}.
$$

Using that $\phi^n = p^n$ since $\beta_q = \beta_1 := 1$ for BDF1, we recast (74.4) as $\boldsymbol{u}^n + \tau\nabla p^n = \widetilde{\boldsymbol{u}}^n$. We square this identity, integrate over D, and use that $\boldsymbol{u}^n$ is divergence-free to obtain

$$
\|\boldsymbol{u}^n\|^2_{\boldsymbol{L}^2} + \tau^2\|\nabla p^n\|^2_{\boldsymbol{L}^2} = \|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{L}^2}.
$$

Summing this identity to the above estimate yields

$$\|\boldsymbol{u}^n\|_{\boldsymbol{L}^2}^2 + \tau^2\|\nabla p^n\|_{\boldsymbol{L}^2}^2 + 2\mu\tau\|\widetilde{\boldsymbol{u}}^n\|_{\boldsymbol{V}}^2 \le \frac{\tau\rho}{2}\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}^2 + \|\boldsymbol{u}^{n-1}\|_{\boldsymbol{L}^2}^2.$$

Summing the result over $n \in \mathcal{N}_\tau$ yields the assertion.

Exercise 74.2 (Curl-div-grad identity). Assume first that $\boldsymbol{v} \in \boldsymbol{C}_0^1(D)$. Using the integration by parts formulae (4.8) in the identity

$$\int_D -\Delta\boldsymbol{v}\cdot\boldsymbol{v}\,\mathrm{d}x = \int_D \big(-\nabla(\nabla\cdot\boldsymbol{v}) + \nabla\times(\nabla\times\boldsymbol{v})\big)\cdot\boldsymbol{v}\,\mathrm{d}x,$$

we obtain

$$\int_D \nabla\boldsymbol{v}{:}\nabla\boldsymbol{v}\,\mathrm{d}x = \int_D (\nabla\cdot\boldsymbol{v})^2\,\mathrm{d}x + \int_D (\nabla\times\boldsymbol{v})^2\,\mathrm{d}x,$$

which is the expected identity. This identity is extended to $\boldsymbol{H}_0^1(D)$ by density.

Exercise 74.3 (Inverse of the Stokes operator). (i) Owing to the definition of $\boldsymbol{S}(\boldsymbol{v})$, we have

$$2\mu\big(\mathbb{e}(\boldsymbol{S}(\boldsymbol{v})),\mathbb{e}(\boldsymbol{v})\big)_{\mathbb{L}^2} = (r,\nabla\cdot\boldsymbol{v})_{L^2} + \|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2,$$

since the duality pairing $\langle\cdot,\cdot\rangle_{\boldsymbol{V}',\boldsymbol{V}}$ is an extension of the $\boldsymbol{L}^2$-inner product. This implies that $2\mu\big(\mathbb{e}(\boldsymbol{S}(\boldsymbol{v})),\mathbb{e}(\boldsymbol{v})\big)_{\mathbb{L}^2} = \|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2$ for all $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}$.
(ii) Assume that $\boldsymbol{v} \in \boldsymbol{V} := \boldsymbol{H}_0^1(D)$. Owing to the definition of $\boldsymbol{S}(\boldsymbol{v})$, we have for all $\boldsymbol{v}^\star \in \boldsymbol{\mathcal{H}}$,

$$\begin{aligned}
2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{v})),\mathbb{e}(\boldsymbol{v}))_{\mathbb{L}^2} &= (r,\nabla\cdot\boldsymbol{v})_{L^2} + \|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2\\
&= (r,\nabla\cdot(\boldsymbol{v}-\boldsymbol{v}^\star))_{L^2} + \|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2\\
&= -(\nabla r,\boldsymbol{v}-\boldsymbol{v}^\star)_{\boldsymbol{L}^2} + \|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2\\
&\ge -|r|_{H^1}\|\boldsymbol{v}-\boldsymbol{v}^\star\|_{\boldsymbol{L}^2} + \|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2\\
&\ge -c\,\|\boldsymbol{v}\|_{\boldsymbol{L}^2}\|\boldsymbol{v}-\boldsymbol{v}^\star\|_{\boldsymbol{L}^2} + \|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2\\
&\ge -c(\gamma)\,\|\boldsymbol{v}-\boldsymbol{v}^\star\|_{\boldsymbol{L}^2}^2 + (1-\gamma)\|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2,
\end{aligned}$$

for all $\gamma \in (0,1)$, where the last bound follows from Young's inequality. This completes the proof of the assertion.
(iii) For all $\boldsymbol{v},\boldsymbol{w} \in \boldsymbol{V}' = \boldsymbol{H}^{-1}(D)$, we have

$$\langle\boldsymbol{w},\boldsymbol{S}(\boldsymbol{v})\rangle_{\boldsymbol{V}',\boldsymbol{V}} = 2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{v})),\mathbb{e}(\boldsymbol{S}(\boldsymbol{w})))_{\mathbb{L}^2} = \langle\boldsymbol{v},\boldsymbol{S}(\boldsymbol{w})\rangle_{\boldsymbol{V}',\boldsymbol{V}}.$$

Hence, the bilinear form $\boldsymbol{V}'\times\boldsymbol{V}' \ni (\boldsymbol{v},\boldsymbol{w}) \mapsto \langle\boldsymbol{w},\boldsymbol{S}(\boldsymbol{v})\rangle_{\boldsymbol{V}',\boldsymbol{V}} \in \mathbb{R}$ is symmetric. This bilinear form is also positive since $\langle\boldsymbol{v},\boldsymbol{S}(\boldsymbol{v})\rangle_{\boldsymbol{V}',\boldsymbol{V}} = 2\mu\|\boldsymbol{S}(\boldsymbol{v})\|_{\boldsymbol{V}}^2$. Hence, the map $\boldsymbol{v} \mapsto |\boldsymbol{v}|_\star := \langle\boldsymbol{v},\boldsymbol{S}(\boldsymbol{v})\rangle_{\boldsymbol{V}',\boldsymbol{V}}^{\frac12}$ induces a seminorm on $\boldsymbol{V}' = \boldsymbol{H}^{-1}(D)$. Notice finally that

$$|\boldsymbol{v}|_\star^2 := \langle\boldsymbol{v},\boldsymbol{S}(\boldsymbol{v})\rangle_{\boldsymbol{V}',\boldsymbol{V}} \le \|\boldsymbol{v}\|_{\boldsymbol{V}'}\|\boldsymbol{S}(\boldsymbol{v})\|_{\boldsymbol{V}}.$$

But $2\mu\|\boldsymbol{S}(\boldsymbol{v})\|_{\boldsymbol{V}}^2 \le \|\boldsymbol{v}\|_{\boldsymbol{V}'}\|\boldsymbol{S}(\boldsymbol{v})\|_{\boldsymbol{V}}$, so that $2\mu\|\boldsymbol{S}(\boldsymbol{v})\|_{\boldsymbol{V}} \le \|\boldsymbol{v}\|_{\boldsymbol{V}'}$. Hence, we have

$$|\boldsymbol{v}|_\star \le (2\mu)^{-\frac12}\|\boldsymbol{v}\|_{\boldsymbol{V}'}.$$

Exercise 74.4 (Lemma 74.11). (i) We write $\boldsymbol{e} := \boldsymbol{u}^\varepsilon - \boldsymbol{u}$ and $q := p^\varepsilon - p$. Subtracting (74.14a) from (75.1), we find

$$\partial_t \boldsymbol{e} - \nabla\cdot\mathbb{s}(\boldsymbol{e}) + \nabla q = \boldsymbol{0}, \qquad \boldsymbol{e}|_{\partial D} = \boldsymbol{0}, \qquad \boldsymbol{e}(0) = \boldsymbol{0}, \tag{74.1a}$$

$$\nabla\cdot\boldsymbol{e} - \varepsilon\Delta\phi^\varepsilon = 0, \qquad \boldsymbol{n}\cdot\nabla\phi^\varepsilon_{|\partial D} = 0, \tag{74.1b}$$

$$\varepsilon\partial_t q = \phi^\varepsilon - \mu\nabla\cdot\boldsymbol{e} - \varepsilon\partial_t p, \qquad q(0) = 0. \tag{74.1c}$$

Taking the inner product of the time derivative of (74.1a) with $\partial_t\boldsymbol{e}$, using the coercivity of the bilinear form $a(\boldsymbol{v},\boldsymbol{w}) := (\mathbb{s}(\boldsymbol{v}),\mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2}$, and integrating by parts the term involving q, we find

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\partial_t\boldsymbol{e}\|_{\boldsymbol{L}^2}^2 + 2\mu\|\partial_t\boldsymbol{e}\|_{\boldsymbol{V}}^2 - (\partial_t q, \nabla\cdot\partial_t\boldsymbol{e})_{L^2} = 0.$$

Taking the time derivative of (74.1b), multiplying the result by $\partial_t q$, and using (74.1c) gives

$$\begin{aligned}
-(\partial_t q, \nabla\cdot\partial_t\boldsymbol{e})_{L^2} &= -\varepsilon(\partial_t q, \Delta\partial_t\phi^\varepsilon)_{L^2} \\
&= -(\phi^\varepsilon - \varepsilon\mu\Delta\phi^\varepsilon - \varepsilon\partial_t p, \Delta\partial_t\phi^\varepsilon)_{L^2} \\
&= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla\phi^\varepsilon\|_{\boldsymbol{L}^2}^2 + \frac{1}{2}\varepsilon\mu\frac{\mathrm{d}}{\mathrm{d}t}\|\Delta\phi^\varepsilon\|_{L^2}^2 - \varepsilon(\nabla\partial_t p, \nabla\partial_t\phi^\varepsilon)_{\boldsymbol{L}^2} \\
&= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla\phi^\varepsilon\|_{\boldsymbol{L}^2}^2 + \frac{1}{2}\varepsilon\mu\frac{\mathrm{d}}{\mathrm{d}t}\|\Delta\phi^\varepsilon\|_{L^2}^2 - \varepsilon\frac{\mathrm{d}}{\mathrm{d}t}(\nabla\partial_t p, \nabla\phi^\varepsilon)_{\boldsymbol{L}^2} + \varepsilon(\nabla\partial_{tt} p, \nabla\phi^\varepsilon)_{\boldsymbol{L}^2}.
\end{aligned}$$

The above two relations lead to

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\partial_t\boldsymbol{e}\|_{\boldsymbol{L}^2}^2 + 2\mu\|\partial_t\boldsymbol{e}\|_{\boldsymbol{V}}^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla\phi^\varepsilon\|_{\boldsymbol{L}^2}^2 + \frac{1}{2}\varepsilon\mu\frac{\mathrm{d}}{\mathrm{d}t}\|\Delta\phi^\varepsilon\|_{L^2}^2 \\
= \varepsilon\frac{\mathrm{d}}{\mathrm{d}t}(\nabla\partial_t p, \nabla\phi^\varepsilon)_{\boldsymbol{L}^2} - \varepsilon(\nabla\partial_{tt} p, \nabla\phi^\varepsilon)_{\boldsymbol{L}^2}.
\end{aligned}$$

(ii) Since $\boldsymbol{e}(0) = \boldsymbol{0}$ and $q(0) = 0$, we also have $\nabla\phi^\varepsilon(0) = \boldsymbol{0}$ (since $\Delta\phi^\varepsilon(0) = 0$ and $\boldsymbol{n}\cdot\nabla\phi^\varepsilon(0)_{|\partial D} = 0$) and $\partial_t\boldsymbol{e}(0) = \nabla\cdot\mathbb{s}(\boldsymbol{e}(0)) - \nabla q(0) = \boldsymbol{0}$. After integrating in time from 0 to t the identity derived in Step (i) for all $t \in J$, we obtain

$$\begin{aligned}
\frac{1}{2}\|\partial_t\boldsymbol{e}(t)\|_{\boldsymbol{L}^2}^2 + \frac{1}{2}\|\nabla\phi^\varepsilon(t)\|_{\boldsymbol{L}^2}^2 + \frac{1}{2}\varepsilon\mu\|\Delta\phi^\varepsilon(t)\|_{L^2}^2 + 2\mu\int_0^t \|\partial_t\boldsymbol{e}(s)\|_{\boldsymbol{V}}^2 ds \\
\le \varepsilon\|\nabla\partial_t p(t)\|_{\boldsymbol{L}^2}\|\nabla\phi^\varepsilon(t)\|_{\boldsymbol{L}^2} + \varepsilon\|\nabla\partial_{tt} p\|_{L^\infty(J,\boldsymbol{L}^2)}\int_0^t \|\nabla\phi^\varepsilon(s)\|_{\boldsymbol{L}^2}\,\mathrm{d}s,
\end{aligned} \tag{74.2}$$

where we used the Cauchy–Schwarz inequality for the first term on the right-hand side and Hölder's inequality for the second term. In particular, (74.2) implies that

$$\frac{1}{4}\|\nabla\phi^\varepsilon(t)\|_{\boldsymbol{L}^2}^2 \le \varepsilon^2\|\nabla\partial_t p(t)\|_{\boldsymbol{L}^2}^2 + \varepsilon\|\nabla\partial_{tt} p\|_{L^\infty(J,\boldsymbol{L}^2)}\int_0^t \|\nabla\phi^\varepsilon(s)\|_{\boldsymbol{L}^2}\,\mathrm{d}s.$$

Invoking the Cauchy–Schwarz inequality in time to bound $\int_0^t \|\nabla\phi^\varepsilon(s)\|_{\boldsymbol{L}^2}\,\mathrm{d}s$, followed by Young's inequality, we infer that

$$\frac{1}{4}\|\nabla\phi^\varepsilon(t)\|_{\boldsymbol{L}^2}^2 \le \varepsilon^2\big(\|\nabla\partial_t p(t)\|_{\boldsymbol{L}^2}^2 + t\|\nabla\partial_{tt} p\|_{L^\infty(J,\boldsymbol{L}^2)}^2\big) + \frac{1}{4}\int_0^t \|\nabla\phi^\varepsilon(s)\|_{\boldsymbol{L}^2}^2\,\mathrm{d}s.$$

An application of Gronwall's lemma (see Exercise 65.3) leads to

$$\|\nabla\phi^\varepsilon(t)\|_{\boldsymbol{L}^2}^2 \le c(p,T)\varepsilon^2,$$

for all $t \in J$. Substituting this bound in the right-hand side of (74.2) yields

$$\|\partial_t \boldsymbol{e}(t)\|_{\boldsymbol{L}^2}^2 + \|\nabla\phi^\varepsilon(t)\|_{\boldsymbol{L}^2}^2 + \varepsilon\mu\|\Delta\phi^\varepsilon(t)\|_{L^2}^2 + 2\mu\int_0^t \|\partial_t \boldsymbol{e}(s)\|_{\boldsymbol{V}}^2 ds \le c(p,T)\varepsilon^2.$$

From the above inequality, we obtain immediately

$$\|\nabla\cdot\boldsymbol{u}^\varepsilon\|_{L^\infty(J;L^2)}^2 = \varepsilon^2\|\Delta\phi^\varepsilon\|_{L^\infty(J;L^2)}^2 \le c(p,T)\mu^{-1}\varepsilon^3.$$

(iii) By definition of the Leray projection $\boldsymbol{P}_{\boldsymbol{\mathcal{H}}}$, we can write $\boldsymbol{e} - \boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{e}) = \nabla r$ with $\boldsymbol{n}\cdot\nabla r_{|\partial D} = 0$. Consequently, we have $\nabla\cdot\boldsymbol{e} = \Delta r$, and from (74.1b), we infer that $r = \varepsilon\phi^\varepsilon$ and

$$\|\boldsymbol{e} - \boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{e})\|_{\boldsymbol{L}^2}^2 = \|\nabla r\|_{\boldsymbol{L}^2}^2 = \varepsilon^2\|\nabla\phi^\varepsilon\|_{\boldsymbol{L}^2}^2.$$

We take the inner product of (74.1a) with $\boldsymbol{S}(\boldsymbol{e})$, where $\boldsymbol{S}$ is the inverse of the Stokes operator defined in Exercise 74.3. Recall that $|\boldsymbol{v}|_\star := (\boldsymbol{v}, \boldsymbol{S}(\boldsymbol{v}))_{\boldsymbol{V}',\boldsymbol{V}}^{\frac12}$ denotes the associated seminorm. Since the $\boldsymbol{L}^2$-inner product is an extension of the duality between $\boldsymbol{V}'$ and $\boldsymbol{V}$ and since $\boldsymbol{S}(\boldsymbol{e}) \in \boldsymbol{\mathcal{V}}$, we obtain

$$\frac12\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{e}|_\star^2 + 2\mu(\mathbb{e}(\boldsymbol{e}), \mathbb{e}(\boldsymbol{S}(\boldsymbol{e})))_{\mathbb{L}^2} = 0. \tag{74.3}$$

In Step (ii) of Exercise 74.3, it is proved that for all $\gamma \in (0,1)$, there exists $c(\gamma)$ such that

$$2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{v})), \mathbb{e}(\boldsymbol{v}))_{\mathbb{L}^2} \ge (1-\gamma)\|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2 - c(\gamma)\|\boldsymbol{v} - \boldsymbol{v}^\star\|_{\boldsymbol{L}^2}^2, \quad \forall \boldsymbol{v}^\star \in \boldsymbol{\mathcal{H}}.$$

Using this bound with $\gamma := \frac12$, $\boldsymbol{v} := \boldsymbol{e}$, and $\boldsymbol{v}^\star := \boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{e})$, and recalling that $\|\boldsymbol{e} - \boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{e})\|_{\boldsymbol{L}^2}^2 = \varepsilon^2\|\nabla\phi^\varepsilon\|_{\boldsymbol{L}^2}^2$, we infer that

$$\begin{aligned}\frac12\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{e}|_\star^2 + \frac12\|\boldsymbol{e}\|_{\boldsymbol{L}^2}^2 &\le \frac12\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{e}|_\star^2 + 2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{e})), \mathbb{e}(\boldsymbol{e}))_{\mathbb{L}^2} + c\,\|\boldsymbol{e} - \boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{e})\|_{\boldsymbol{L}^2}^2 \\ &= c\,\|\boldsymbol{e} - \boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{e})\|_{\boldsymbol{L}^2}^2 = c\,\varepsilon^2\|\nabla\phi^\varepsilon\|_{\boldsymbol{L}^2}^2.\end{aligned}$$

This bound holds for all $t \in J$, and we have shown in Step (ii) that $\|\nabla\phi^\varepsilon(t)\|_{\boldsymbol{L}^2}^2 \le c(p,T)\varepsilon^2$ for all $t \in J$. This implies that $\frac12\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{e}|_\star^2 + \frac12\|\boldsymbol{e}\|_{\boldsymbol{L}^2}^2 \le c(p,T)\varepsilon^4$. Integrating this inequality in time over J, we infer that

$$|\boldsymbol{e}(T)|_\star^2 + \int_J \|\boldsymbol{e}(s)\|_{\boldsymbol{L}^2}^2 ds \le c(p,T)\varepsilon^4.$$

This shows that $\|\boldsymbol{e}\|_{L^2(J;\boldsymbol{L}^2)} = \|\boldsymbol{u} - \boldsymbol{u}^\varepsilon\|_{L^2(J;\boldsymbol{L}^2)} \le c(p,T)\varepsilon^2$.

Exercise 74.5 (Gauge-Uzawa). (i) We write the pressure-correction algorithm in rotational form using BDF1 and $p^{\star,n} := p^{n-1}$. We set $\boldsymbol{u}^0 := \boldsymbol{u}_0$, $p^0 := p(0)$, then solve for all $n \in \mathcal{N}_\tau$,

$$\begin{aligned}&\frac{\widetilde{\boldsymbol{u}}^n - \boldsymbol{u}^{n-1}}{\tau} - \mu\Delta\widetilde{\boldsymbol{u}}^n + \nabla p^{n-1} = \boldsymbol{f}^n, \quad \widetilde{\boldsymbol{u}}_{|\partial D}^n = \boldsymbol{0},\\ &\boldsymbol{u}^n + \tau\nabla\phi^n = \widetilde{\boldsymbol{u}}^n, \quad \nabla\cdot\boldsymbol{u}^n = 0, \quad \boldsymbol{u}_{|\partial D}^n\cdot\boldsymbol{n} = 0,\\ &p^n = p^{n-1} + \phi^n - \mu\nabla\cdot\widetilde{\boldsymbol{u}}^n.\end{aligned}$$

(ii) The momentum equations in both algorithms are identical if we set $\boldsymbol{v}_\tau := \boldsymbol{u}_\tau$ and $q_\tau := p_\tau$. The projection step in the gauge-Uzawa technique reduces to the projection step in the pressure-correction method by setting $\phi^n := \psi^n - \psi^{n-1} = \tau(\delta_\tau\psi_\tau)^n$. Finally, the gauge-Uzawa technique gives

$$q^n = q^{n-1} + \psi^n - \psi^{n-1} + \mu(r^n - r^{n-1}) = q^{n-1} + \phi^n - \mu\nabla\cdot\widetilde{\boldsymbol{u}}^n,$$

so that we recover the pressure update in the pressure-correction algorithm in rotational form. Hence, up to the appropriate change of variables, the two algorithms are identical.
(iii) Testing the momentum equation in the gauge-Uzawa algorithm with $2\tau\widetilde{\boldsymbol{v}}^n$ yields

$$\|\widetilde{\boldsymbol{v}}^n\|_{\boldsymbol{L}^2}^2 + \|\widetilde{\boldsymbol{v}}^n - \boldsymbol{v}^{n-1}\|_{\boldsymbol{L}^2}^2 + 2\mu\tau\|\nabla\widetilde{\boldsymbol{v}}^n\|_{\mathbb{L}^2}^2 - 2\tau(\nabla\cdot\widetilde{\boldsymbol{v}}^n, q^{n-1})_{L^2} \\ \leq \|\boldsymbol{v}^{n-1}\|_{\boldsymbol{L}^2}^2 + 2\tau\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}\|\widetilde{\boldsymbol{v}}^n\|_{\boldsymbol{L}^2}.$$

We square the second equation, we square the third one and multiply the result by $\mu\tau$. This gives

$$\|\boldsymbol{v}^n\|_{\boldsymbol{L}^2}^2 + \tau^2\|\nabla\psi^n\|_{\boldsymbol{L}^2}^2 = \|\widetilde{\boldsymbol{v}}^n\|_{\boldsymbol{L}^2}^2 - 2\tau(\nabla\cdot\widetilde{\boldsymbol{v}}^n, \psi^{n-1})_{L^2} + \tau^2\|\nabla\psi^{n-1}\|_{\boldsymbol{L}^2}^2,$$
$$\mu\tau\|r^n\|_{L^2}^2 = \mu\tau\|r^{n-1}\|_{L^2}^2 + \mu\tau\|\nabla\cdot\widetilde{\boldsymbol{v}}^n\|_{L^2}^2 - 2\mu\tau(\nabla\cdot\widetilde{\boldsymbol{v}}^n, r^{n-1})_{L^2}.$$

We add the first inequality to the above two identities and obtain

$$\begin{aligned}&\|\boldsymbol{v}^n\|_{\boldsymbol{L}^2}^2 + \tau^2\|\nabla\psi^n\|_{\boldsymbol{L}^2}^2 + \mu\tau\|r^n\|_{L^2}^2 + \|\widetilde{\boldsymbol{v}}^n - \boldsymbol{v}^{n-1}\|_{\boldsymbol{L}^2}^2 + 2\mu\tau\|\nabla\widetilde{\boldsymbol{v}}^n\|_{\mathbb{L}^2}^2 \\ &\leq \|\boldsymbol{v}^{n-1}\|_{\boldsymbol{L}^2}^2 + \tau^2\|\nabla\psi^{n-1}\|_{\boldsymbol{L}^2}^2 + \mu\tau\|r^{n-1}\|_{L^2}^2 + \rho\tau\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}^2 \\ &\quad + \frac{1}{2}\mu\tau\|\nabla\widetilde{\boldsymbol{v}}^n\|_{\mathbb{L}^2}^2 + \mu\tau\|\nabla\cdot\widetilde{\boldsymbol{v}}^n\|_{L^2}^2 + 2\tau(\nabla\cdot\widetilde{\boldsymbol{v}}^n, q^{n-1} - \psi^{n-1} - \mu r^{n-1})_{L^2},\end{aligned}$$

where we used the Poincaré–Steklov inequality and Young's inequality to infer that

$$2\tau\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}\|\widetilde{\boldsymbol{v}}^n\|_{\boldsymbol{L}^2} \leq 2\tau\frac{\ell_D}{C_{\mathrm{PS}}}\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}\|\nabla\widetilde{\boldsymbol{v}}^n\|_{\mathbb{L}^2} \leq \rho\tau\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}^2 + \frac{1}{2}\mu\tau\|\nabla\widetilde{\boldsymbol{v}}^n\|_{\mathbb{L}^2}^2.$$

We now use the equation $q^{n-1} = \psi^{n-1} + \mu r^{n-1}$ for all $n \in \mathcal{N}_\tau$, $n \geq 2$, and notice that this equation also holds true for $n = 1$ since the initialization enforces $q^0 = \psi^0$ and $r^0 = 0$. Using also that $\|\nabla\cdot\widetilde{\boldsymbol{v}}^n\|_{L^2} \leq \|\nabla\widetilde{\boldsymbol{v}}^n\|_{\mathbb{L}^2}$ (see Exercise 74.2), we obtain

$$\|\boldsymbol{v}^n\|_{\boldsymbol{L}^2}^2 + \tau^2\|\nabla\psi^n\|_{\boldsymbol{L}^2}^2 + \mu\tau\|r^n\|_{L^2}^2 + \|\widetilde{\boldsymbol{v}}^n - \boldsymbol{v}^{n-1}\|_{\boldsymbol{L}^2}^2 + \frac{1}{2}\mu\tau\|\nabla\widetilde{\boldsymbol{v}}^n\|_{\mathbb{L}^2}^2 \\ \leq \|\boldsymbol{v}^{n-1}\|_{\boldsymbol{L}^2}^2 + \tau^2\|\nabla\psi^{n-1}\|_{\boldsymbol{L}^2}^2 + \mu\tau\|r^{n-1}\|_{L^2}^2 + \rho\tau\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}^2.$$

Chapter 75

Artificial compressibility

Exercises

Exercise 75.1 (Lemma 75.1). (i) Prove (75.4). (*Hint*: test the momentum equation with $\boldsymbol{v}$ and the mass equation with q, use Lemma 53.9 to bound $(q,g)_{L^2}$, integrate in time from 0 to t for all $t \in J$, and integrate by parts in time.) (ii) Prove (75.5). (*Hint*: use the inf-sup condition on the bilinear form b together with the bounds derived in Step (i).)

Exercise 75.2 (Lemma 75.2). (i) Let $\delta_\tau \boldsymbol{k}^n := \frac{\boldsymbol{k}^n - \boldsymbol{k}^{n-1}}{\tau}$ and $\delta_\tau g^n := \frac{g^n - g^{n-1}}{\tau}$ for all $n \in \mathcal{N}_\tau$. Prove that $\|\delta_\tau \boldsymbol{k}_\tau\|_{\ell^2(J_*;\boldsymbol{L}^2)} \le \|\partial_t \boldsymbol{k}\|_{L^2(J_*;\boldsymbol{L}^2)}$. Let $\Gamma(t) := \frac{1}{\tau}\int_{t-\tau}^t \partial_\xi g(\xi)\,\mathrm{d}\xi$ for all $t \in J_*$. Prove that $\partial_t \Gamma(t) = \frac{1}{\tau}\int_{t-\tau}^t \partial_{\xi\xi} g(\xi)\,\mathrm{d}\xi$ for all $t \in J_*$ and that $\|\partial_t \Gamma\|_{L^2(J_{**},L^2)} \le \|\partial_{\xi\xi} g\|_{L^2(J_*;L^2)}$. (*Hint*: use the Cauchy–Schwarz inequality and Fubini's theorem.) (ii) Derive the system satisfied by the time sequences $\delta_\tau \boldsymbol{u}_\tau := (\frac{\boldsymbol{u}^n - \boldsymbol{u}^{n-1}}{\tau})_{n\in\mathcal{N}_\tau}$ and $\delta_\tau p_\tau := (\frac{p^n - p^{n-1}}{\tau})_{n\in\mathcal{N}_\tau}$. (iii) Prove the estimate (75.10). (*Hint*: use the inf-sup condition on the bilinear form b and bound $\delta_\tau \boldsymbol{u}_\tau$ by adapting the proof of (75.9).)

Exercise 75.3 (Proposition 75.3). The goal of this exercise is to prove Proposition 75.3. (i) Let $\boldsymbol{e}_\tau := \boldsymbol{u}_\tau - \boldsymbol{\pi}_\tau(\boldsymbol{u})$ and $r_\tau := p_\tau - \pi_\tau(p)$. Let $\boldsymbol{\psi}(t) := \frac{1}{\tau}\int_{t-\tau}^t (\xi - t + \tau)\partial_{\xi\xi}\boldsymbol{u}\,\mathrm{d}\xi$ and $\phi(t) := -\frac{1}{\lambda}\int_{t-\tau}^t \partial_\xi p\,\mathrm{d}\xi$ for all $t \in J_*$. Show that

$$\begin{cases} \dfrac{1}{\tau}(\boldsymbol{e}^n - \boldsymbol{e}^{n-1}) - \nabla\cdot\mathbb{s}(\boldsymbol{e}^n) + \nabla r^n = \boldsymbol{\psi}^n, \quad \boldsymbol{e}^n_{|\partial D} = \boldsymbol{0}, \\ \dfrac{1}{\lambda}(r^n - r^{n-1}) + \nabla\cdot\boldsymbol{e}^n = \phi^n. \end{cases}$$

(ii) Prove the estimates (75.11) and (75.12). (*Hint*: use Lemma 75.2.)

Exercise 75.4 (Initialization). Let $\boldsymbol{u}_0$ be the initial velocity, and assume that $p(0)$ is given. Let $t_1 := \tau$. Using the first-order artificial compressibility algorithm (75.6) and Richardson's extrapolation, propose a technique to estimate $(\partial_{tt}\boldsymbol{u}(t_1), \partial_{tt}p(t_1))$ with $\mathcal{O}(\tau)$ accuracy, $(\partial_t\boldsymbol{u}(t_1), \partial_t p(t_1))$ with $\mathcal{O}(\tau^2)$ accuracy, and $(\boldsymbol{u}(t_1), p(t_1))$ with $\mathcal{O}(\tau^3)$ accuracy. (*Hint*: estimate $(\boldsymbol{u}, p)$ at the times t_1 and $t_2 := 2\tau$ by using (75.6) with the time steps $\frac{\tau}{3}$, $\frac{\tau}{2}$, and τ, keeping λ fixed. Conclude by using finite differences centered at $t_1 := \tau$.)

Solution to exercises

Exercise 75.1 (Lemma 75.1). (i) Testing the momentum equation with $\boldsymbol{v}$ and the mass equation with q, adding the two results, using the coercivity of the bilinear form $a(\boldsymbol{v},\boldsymbol{w}) := (\mathbb{s}(\boldsymbol{v}),\mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2}$ and Young's inequality to estimate $\langle \boldsymbol{k},\boldsymbol{v}\rangle_{\boldsymbol{V}',\boldsymbol{V}}$ gives

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2+\frac{1}{2}\epsilon\frac{\mathrm{d}}{\mathrm{d}t}\|q\|_{L^2}^2+\mu\|\boldsymbol{v}\|_{\boldsymbol{V}}^2\le\frac{1}{4\mu}\|\boldsymbol{k}\|_{\boldsymbol{V}'}^2+(q,g)_{L^2}. \tag{75.1}$$

Owing to Lemma 53.9, there exists β_D such that for all $g\in L^2_*(D)$, there is $\boldsymbol{w}(g)\in\boldsymbol{V}$ s.t. $\nabla\cdot(\boldsymbol{w}(g))=g$ and $\beta_D\|\boldsymbol{w}(g)\|_{\boldsymbol{V}}\le\|g\|_{L^2(D)}$. We infer that

$$\begin{aligned}(q,g)_{L^2}&=(q,\nabla\cdot(\boldsymbol{w}(g))_{L^2}\\&=-(\nabla q,\boldsymbol{w}(g))_{\boldsymbol{L}^2}\\&=(-\boldsymbol{k}+\partial_t\boldsymbol{v}-\nabla\cdot\mathbb{s}(\boldsymbol{v}),\boldsymbol{w}(g))_{\boldsymbol{L}^2}\\&\le\|\boldsymbol{k}\|_{\boldsymbol{V}'}\|\boldsymbol{w}(g)\|_{\boldsymbol{V}}+2\mu\|\boldsymbol{v}\|_{\boldsymbol{V}}\|\boldsymbol{w}(g)\|_{\boldsymbol{V}}-(\partial_t\boldsymbol{v},\boldsymbol{w}(g))_{\boldsymbol{L}^2}.\end{aligned} \tag{75.2}$$

We integrate the above inequality over time from 0 to t for all $t\in J$, and set $J_t:=(0,t)$. Let us first consider the third term on the right-hand side. Integrating by parts in time and using the property $\partial_t\boldsymbol{w}(g)=\boldsymbol{w}(\partial_t g)$, we obtain

$$\begin{aligned}&\int_0^t(\partial_t\boldsymbol{v},\boldsymbol{w}(g))_{\boldsymbol{L}^2}\,\mathrm{d}s\\&=-\int_0^t(\boldsymbol{v},\boldsymbol{w}(\partial_t g))_{\boldsymbol{L}^2}\,\mathrm{d}s+(\boldsymbol{v}(t),\boldsymbol{w}(g(t)))_{\boldsymbol{L}^2}-(\boldsymbol{v}(0),\boldsymbol{w}(g(0)))_{\boldsymbol{L}^2}\\&\le\|\boldsymbol{v}\|_{L^2(J_t;\boldsymbol{L}^2)}\|\boldsymbol{w}(\partial_t g)\|_{L^2(J_t;\boldsymbol{L}^2)}+\|\boldsymbol{v}(t)\|_{\boldsymbol{L}^2}\|\boldsymbol{w}(g(t))\|_{\boldsymbol{L}^2}+\|\boldsymbol{v}(0)\|_{\boldsymbol{L}^2}\|\boldsymbol{w}(g(0))\|_{\boldsymbol{L}^2}\\&\le\beta_D^{-1}\Big(\tfrac{\ell_D^2}{C_{\mathrm{KPS}}^2}\|\boldsymbol{v}\|_{L^2(J_t;\boldsymbol{V})}\|\partial_t g\|_{L^2(J_t;L^2)}+\|\boldsymbol{v}(t)\|_{\boldsymbol{L}^2}\tfrac{\ell_D}{C_{\mathrm{KPS}}}\|g(t)\|_{L^2}+\|\boldsymbol{v}_0\|_{\boldsymbol{L}^2}\tfrac{\ell_D}{C_{\mathrm{KPS}}}\|g_0\|_{L^2}\Big),\end{aligned}$$

where we used the Cauchy–Schwarz inequality, the bound $C_{\mathrm{KPS}}\|\boldsymbol{v}\|_{\boldsymbol{L}^2}\le\ell_D\|\boldsymbol{v}\|_{\boldsymbol{V}}$ for all $\boldsymbol{v}\in\boldsymbol{V}$, and the above bound on the lifting $\boldsymbol{w}$. Using this bound in (75.2) yields

$$\begin{aligned}\int_0^t(q,g)_{L^2}\,\mathrm{d}s\le{}&\beta_D^{-1}(\|\boldsymbol{k}\|_{L^2(J_t;\boldsymbol{V}')}+2\mu\|\boldsymbol{v}\|_{L^2(J_t;\boldsymbol{V})})\|g\|_{L^2(J_t;L^2)}\\&+\beta_D^{-1}\Big(\tfrac{\ell_D^2}{C_{\mathrm{KPS}}^2}\|\boldsymbol{v}\|_{L^2(J_t;\boldsymbol{V})}\|\partial_t g\|_{L^2(J_t;L^2)}+\|\boldsymbol{v}(t)\|_{\boldsymbol{L}^2}\tfrac{\ell_D}{C_{\mathrm{KPS}}}\|g(t)\|_{L^2}\\&+\|\boldsymbol{v}_0\|_{\boldsymbol{L}^2}\tfrac{\ell_D}{C_{\mathrm{KPS}}}\|g_0\|_{L^2}\Big).\end{aligned}$$

Integrating (75.1) from 0 to t, using this bound and the time scale $\rho:=C_{\mathrm{KPS}}^{-2}\frac{\ell_D^2}{\mu}$, we obtain

$$\begin{aligned}&\frac{1}{2}\|\boldsymbol{v}(t)\|_{\boldsymbol{L}^2}^2+\frac{1}{2}\epsilon\|q(t)\|_{L^2}^2+\mu\|\boldsymbol{v}\|_{L^2(J_t;\boldsymbol{V})}^2\\&\le\frac{1}{2}\|\boldsymbol{v}_0\|_{\boldsymbol{L}^2}^2+\frac{1}{2}\epsilon\|q_0\|_{L^2}^2+\frac{1}{4\mu}\|\boldsymbol{k}\|_{L^2(J_t;\boldsymbol{V}')}^2\\&\quad+\beta_D^{-1}(\mu^{-\frac12}\|\boldsymbol{k}\|_{L^2(J_t;\boldsymbol{V}')}+2\mu^{\frac12}\|\boldsymbol{v}\|_{L^2(J_t;\boldsymbol{V})})\mu^{\frac12}\|g\|_{L^2(J_t;L^2)}\\&\quad+\beta_D^{-1}\Big(\mu^{\frac12}\|\boldsymbol{v}\|_{L^2(J_t;\boldsymbol{V})}\mu^{\frac12}\rho\|\partial_t g\|_{L^2(J_t;L^2)}+\|\boldsymbol{v}(t)\|_{\boldsymbol{L}^2}\mu^{\frac12}\rho^{\frac12}\|g(t)\|_{L^2}+\|\boldsymbol{v}_0\|_{\boldsymbol{L}^2}\mu^{\frac12}\rho^{\frac12}\|g_0\|_{L^2}\Big).\end{aligned}$$

Using Young inequalities and since $\|g\|^2_{H^1(J;L^2)} = \|g\|^2_{L^2(J;L^2)} + \rho^2\|\partial_t g\|^2_{L^2(J;L^2)}$, we obtain

$$\begin{aligned}&\frac{1}{4}\|\boldsymbol{v}(t)\|^2_{\boldsymbol{L}^2} + \frac{1}{2}\epsilon\|q(t)\|^2_{L^2} + \frac{1}{2}\mu\|\boldsymbol{v}\|^2_{L^2(J_t;\boldsymbol{V})}\\ &\le \|\boldsymbol{v}_0\|^2_{\boldsymbol{L}^2} + \frac{1}{2}\epsilon\|q_0\|^2_{L^2} + c\big(\mu^{-1}\|\boldsymbol{k}\|^2_{L^2(J;\boldsymbol{V}')} + \mu\|g\|^2_{H^1(J;L^2)}\big),\end{aligned}$$

where we used that $\|\boldsymbol{k}\|_{L^2(J_t;\boldsymbol{V}')} \le \|\boldsymbol{k}\|_{L^2(J;\boldsymbol{V}')}$, $\|g\|_{H^1(J_t;L^2)} \le \|g\|_{H^1(J;L^2)}$, and that $\rho^{\frac12}\|g_0\|_{L^2} \le \rho^{\frac12}\|g\|_{C^0(\overline{J};L^2)} \le c\|g\|_{H^1(J;L^2)}$. Taking the supremum over $t \in J$ and then taking the inequality for $t := T$ leads to

$$\begin{aligned}&\frac{1}{4}\|\boldsymbol{v}\|^2_{L^\infty(J;\boldsymbol{L}^2)} + \frac{1}{2}\epsilon\|q\|^2_{L^\infty(J;L^2)} + \frac{1}{2}\mu\|\boldsymbol{v}\|^2_{L^2(J;\boldsymbol{V})}\\ &\le 2\|\boldsymbol{v}_0\|^2_{\boldsymbol{L}^2} + \epsilon\|q_0\|^2_{L^2} + c\big(\mu^{-1}\|\boldsymbol{k}\|^2_{L^2(J;\boldsymbol{V}')} + \mu\|g\|^2_{H^1(J;L^2)}\big).\end{aligned}$$

This proves the estimate (75.4).
(v) Using the inf-sup condition on the bilinear form b, we infer that

$$\begin{aligned}\beta_D\|q\|_{L^2} &\le \sup_{\boldsymbol{w}\in\boldsymbol{V}} \frac{|b(\boldsymbol{w},q)|}{\|\boldsymbol{w}\|_{\boldsymbol{V}}}\\ &= \sup_{\boldsymbol{w}\in\boldsymbol{V}} \frac{|(\partial_t\boldsymbol{v},\boldsymbol{w})_{\boldsymbol{L}^2} + (\mathbb{s}(\boldsymbol{v}),\mathbb{e}(\boldsymbol{w})_{\mathbb{L}^2} - \langle\boldsymbol{k},\boldsymbol{w}\rangle_{\boldsymbol{V}',\boldsymbol{V}}|}{\|\boldsymbol{w}\|_{\boldsymbol{V}}}\\ &\le \frac{\ell_D}{C_{\mathrm{KPS}}}\|\partial_t\boldsymbol{v}\|_{\boldsymbol{L}^2} + 2\mu\|\boldsymbol{v}\|_{\boldsymbol{V}} + \|\boldsymbol{k}\|_{\boldsymbol{V}'}.\end{aligned}$$

Taking the time derivative of the system (75.3) and using the estimate (75.4) for the time derivatives, we infer that

$$\begin{aligned}\frac{1}{\rho}\|\partial_t\boldsymbol{v}\|^2_{L^2(J;\boldsymbol{L}^2)} = \mu\frac{C^2_{\mathrm{KPS}}}{\ell_D^2}\|\partial_t\boldsymbol{v}\|^2_{L^2(J;\boldsymbol{L}^2)} &\le \mu\|\partial_t\boldsymbol{v}\|^2_{L^2(J;\boldsymbol{V})}\\ &\le 4\|\partial_t\boldsymbol{v}(0)\|^2_{\boldsymbol{L}^2} + 2\epsilon\|\partial_t q(0)\|^2_{L^2} + c\left(\mu^{-1}\|\partial_t\boldsymbol{k}\|^2_{L^2(J;\boldsymbol{V}')} + \mu\|\partial_t g\|^2_{H^1(J;L^2)}\right).\end{aligned}$$

This bound together with the estimate on $\|\boldsymbol{v}\|_{L^2(J;\boldsymbol{V})}$ gives

$$\begin{aligned}\beta_D^2\|q\|^2_{L^2(J;L^2)} &\le 3\big(\mu\rho\|\partial_t\boldsymbol{v}\|^2_{L^2(J;\boldsymbol{L}^2)} + 4\mu^2\|\boldsymbol{v}\|^2_{L^2(J;\boldsymbol{V})} + \|\boldsymbol{k}\|^2_{L^2(J;\boldsymbol{V}')}\big)\\ &\le c\Big(\mu\rho^2\big(\|\partial_t\boldsymbol{v}(0)\|^2_{\boldsymbol{L}^2} + \epsilon\|\partial_t q(0)\|^2_{L^2} + \mu^{-1}\|\partial_t\boldsymbol{k}\|^2_{L^2(J;\boldsymbol{V}')} + \mu\|\partial_t g\|^2_{H^1(J;L^2)}\big)\\ &\quad + \mu\big(\|\boldsymbol{v}_0\|^2_{\boldsymbol{L}^2} + \epsilon\|q_0\|^2_{L^2} + \|\boldsymbol{k}\|^2_{L^2(J;\boldsymbol{V}')} + \mu^2\|g\|^2_{H^1(J;L^2)}\big)\Big).\end{aligned}$$

Using that

$$\begin{aligned}\|g\|^2_{H^2(J;L^2)} &= \|g\|^2_{L^2(J;L^2)} + \rho^2\|\partial_t g\|^2_{L^2(J;L^2)} + \rho^4\|\partial_{tt} g\|^2_{L^2(J;L^2)}\\ &\le \|g\|^2_{H^1(J;L^2)} + \rho^2\|\partial_t g\|^2_{H^1(J;L^2)},\end{aligned}$$

and $\|\boldsymbol{k}\|^2_{H^1(J;\boldsymbol{V}')} = \|\boldsymbol{k}\|^2_{L^2(J;\boldsymbol{V}')} + \rho^2\|\partial_t\boldsymbol{k}\|^2_{L^2(J;\boldsymbol{V}')}$, and rearranging the terms proves (75.5).

Exercise 75.2 (Lemma 75.2). (i) We observe that

$$\begin{aligned}\|\delta_\tau\boldsymbol{k}_\tau\|^2_{\ell^2(J_*;\boldsymbol{L}^2)} = \sum_{l\in\{2:N\}}\tau^{-1}\|\boldsymbol{k}^l - \boldsymbol{k}^{l-1}\|^2_{\boldsymbol{L}^2} &= \sum_{l\in\{2:N\}}\tau^{-1}\Big\|\int_{J_l}\partial_t\boldsymbol{k}\mathrm{d}t\Big\|^2_{\boldsymbol{L}^2}\\ &\le \sum_{l\in\{2:N\}}\int_{J_l}\|\partial_t\boldsymbol{k}\|^2_{\boldsymbol{L}^2}\mathrm{d}t = \|\partial_t\boldsymbol{k}\|^2_{L^2(J_*;\boldsymbol{L}^2)}.\end{aligned}$$

Moreover, the function $\Gamma(t) := \frac{1}{\tau}\int_{t-\tau}^{t} \partial_\xi g(\xi)\,\mathrm{d}\xi$ satisfies

$$\partial_t \Gamma(t) = \frac{1}{\tau}(\partial_t g(t) - \partial_t g(t-\tau)) = \frac{1}{\tau}\int_{t-\tau}^{t} \partial_{\xi\xi} g(\xi)\,\mathrm{d}\xi,$$

so that we have

$$\begin{aligned}
\|\partial_t \Gamma\|^2_{L^2(J_{**},L^2)} &= \int_{J_{**}} \frac{1}{\tau^2}\Big\| \int_{t-\tau}^{t} \partial_{\xi\xi} g(\xi)\,\mathrm{d}\xi\Big\|^2_{L^2}\mathrm{d}t \\
&\le \int_{J_{**}} \frac{1}{\tau^2}\Big(\int_{t-\tau}^{t} \|\partial_{\xi\xi} g(\xi)\|_{L^2}\,\mathrm{d}\xi\Big)^2\mathrm{d}t \\
&\le \int_{J_{**}} \frac{1}{\tau}\int_{t-\tau}^{t} \|\partial_{\xi\xi} g(\xi)\|^2_{L^2}\,\mathrm{d}\xi\mathrm{d}t \\
&\le \int_{J_*} \|\partial_{\xi\xi} g(\xi)\|^2_{L^2}\frac{1}{\tau}\int_{t_-(\xi)}^{t_+(\xi)} \mathrm{d}t\,\mathrm{d}\xi \le \|\partial_{\xi\xi} g\|^2_{L^2(J_*;L^2)},
\end{aligned}$$

where we used the Cauchy–Schwarz inequality, Fubini's theorem with $t_-(\xi) := \max(t_2,\xi)$ and $t_+(\xi) := \min(T,\xi+\tau)$ so that $|t_+(\xi) - t_-(\xi)| \le \tau$ for all $\xi \in J_*$.
(ii) By linearity, we observe that the following holds true for all $n \in \mathcal{N}_\tau$, $n \ge 2$:

$$\begin{cases}
\dfrac{1}{\tau}(\delta_\tau \boldsymbol{u}^n - \delta_\tau \boldsymbol{u}^{n-1}) - \nabla\cdot\mathbb{s}(\delta_\tau \boldsymbol{u}^n) + \nabla \delta_\tau p^n = \delta_\tau \boldsymbol{k}^n, \quad \delta_\tau \boldsymbol{u}^n_{|\partial D} = \boldsymbol{0}, \\
\dfrac{1}{\lambda}(\delta_\tau p^n - \delta_\tau p^{n-1}) + \nabla\cdot\delta_\tau \boldsymbol{u}^n = \delta_\tau g^n.
\end{cases}$$

Hence, the pair $(\delta_\tau \boldsymbol{u}_\tau, \delta_\tau p_\tau)$ solves the same system as the pair $(\boldsymbol{u}_\tau, p_\tau)$, except that now the source terms $(\boldsymbol{k}_\tau, g_\tau)$ are replaced by $(\delta_\tau \boldsymbol{k}_\tau, \delta_\tau g_\tau)$.
(iii) Using the inf-sup condition on the bilinear form b, we infer that

$$\begin{aligned}
\beta_D \|p^n\|_{L^2} &\le \sup_{\boldsymbol{w}\in\boldsymbol{V}} \frac{|b(\boldsymbol{w}, p^n)|}{\|\boldsymbol{w}\|_{\boldsymbol{V}}} \\
&= \sup_{\boldsymbol{w}\in\boldsymbol{V}} \frac{|(\frac{\boldsymbol{u}^n - \boldsymbol{u}^{n-1}}{\tau}, \boldsymbol{w})_{\boldsymbol{L}^2} + (\mathbb{s}(\boldsymbol{u}^n), \mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2} - (\boldsymbol{k}^n, \boldsymbol{w})_{\boldsymbol{L}^2}|}{\|\boldsymbol{w}\|_{\boldsymbol{V}}} \\
&\le (\rho\mu)^{\frac12}\|\tfrac{\boldsymbol{u}^n - \boldsymbol{u}^{n-1}}{\tau}\|_{\boldsymbol{L}^2} + 2\mu\|\boldsymbol{u}^n\|_{\boldsymbol{V}} + (\rho\mu)^{\frac12}\|\boldsymbol{k}^n\|_{\boldsymbol{L}^2}.
\end{aligned}$$

This estimate implies that

$$\|p_\tau\|^2_{\ell^2(J;L^2)} \le c\,\big(\rho\mu\|\delta_\tau \boldsymbol{u}_\tau\|^2_{\ell^2(J;\boldsymbol{L}^2)} + \mu^2\|\boldsymbol{u}_\tau\|^2_{\ell^2(J;\boldsymbol{V})} + \rho\mu\|\boldsymbol{k}_\tau\|^2_{\ell^2(J;\boldsymbol{L}^2)}\big). \tag{75.3}$$

It remains to estimate $\rho\mu\|\delta_\tau \boldsymbol{u}_\tau\|^2_{\ell^2(J;\boldsymbol{L}^2)}$ since a bound on $\mu^2\|\boldsymbol{u}_\tau\|^2_{\ell^2(J;\boldsymbol{V})}$ follows by multiplying (75.9) by μ. Owing to Step (ii), we bound $\delta_\tau \boldsymbol{u}_\tau$ by proceeding as in the proof of the estimate (75.9). Recalling the notation $J_* := (t_1, T)$, $J_{**} := (t_2, T)$, observing that $\Gamma(t_n) = \delta_\tau g^n$ for all $n \in \mathcal{N}_\tau$, and using the bounds derived in Step (i), we infer that

$$\begin{aligned}
\|\delta_\tau \boldsymbol{u}_\tau\|^2_{\ell^\infty(J_*;\boldsymbol{L}^2)} + \mu\|\delta_\tau \boldsymbol{u}_\tau\|^2_{\ell^2(J_*;\boldsymbol{V})} \le c\,e^{\frac{4T}{\rho}}\Big(&\|\delta_\tau \boldsymbol{u}^1\|^2_{\boldsymbol{L}^2} + \frac{\tau}{\mu}\|\delta_\tau p^1\|^2_{L^2} + \rho\|\partial_t \boldsymbol{k}\|^2_{L^2(J_*;\boldsymbol{L}^2)} \\
&+ \mu(T+\rho)\|\delta_\tau g_\tau\|^2_{\ell^\infty(J_*;L^2)} + \mu\rho^2\|\partial_{tt} g\|^2_{L^2(J_*;L^2)}\Big).
\end{aligned}$$

As a result, we have

$$\begin{aligned}\rho\mu\|\delta_\tau \boldsymbol{u}_\tau\|^2_{\ell^2(J;\boldsymbol{L}^2)} &= \rho\mu\tau\|\delta_\tau \boldsymbol{u}^1\|^2_{\boldsymbol{L}^2} + \rho\mu\|\delta_\tau \boldsymbol{u}_\tau\|^2_{\ell^2(J_*;\boldsymbol{L}^2)}\\ &\le \rho\mu\tau\|\delta_\tau \boldsymbol{u}^1\|^2_{\boldsymbol{L}^2} + \rho^2\mu^2\|\delta_\tau \boldsymbol{u}_\tau\|^2_{\ell^2(J_*;\boldsymbol{V})}\\ &\le \rho\mu\tau\|\delta_\tau \boldsymbol{u}^1\|^2_{\boldsymbol{L}^2} + c\,e^{\frac{4T}{\rho}}\Big(\rho^2\mu\|\delta_\tau \boldsymbol{u}^1\|^2_{\boldsymbol{L}^2} + \rho^2\tau\|\delta_\tau p^1\|^2_{L^2} + \rho^3\mu\|\partial_t \boldsymbol{k}\|^2_{L^2(J_*;\boldsymbol{L}^2)} +\\ &\quad + \rho^2\mu^2(T+\rho)\|\delta_\tau g_\tau\|^2_{\ell^\infty(J_*;L^2)} + \mu^2\rho^4\|\partial_{tt} g\|^2_{L^2(J_*;L^2)}\Big),\end{aligned}$$

and the term $\rho\mu\tau\|\delta_\tau \boldsymbol{u}^1\|^2_{\boldsymbol{L}^2}$ can be combined with the term $\rho^2\mu\|\delta_\tau \boldsymbol{u}^1\|^2_{\boldsymbol{L}^2}$ since we assumed $\tau \le \frac{1}{4}\rho$. Finally, the estimate on the pressure follows by combining the above bound with (75.3).

Exercise 75.3 (Proposition 75.3). (i) We have

$$\begin{cases}\dfrac{1}{\tau}(\boldsymbol{e}^n - \boldsymbol{e}^{n-1}) - \nabla\cdot\mathbb{s}(\boldsymbol{e}^n) + \nabla r^n = \boldsymbol{\psi}^n, \quad \boldsymbol{e}^n_{|\partial D} = \boldsymbol{0},\\ \dfrac{1}{\lambda}(r^n - r^{n-1}) + \nabla\cdot\boldsymbol{e}^n = \phi^n,\end{cases}$$

where $\boldsymbol{\psi}^n := \boldsymbol{\psi}(t_n)$ and $\phi^n := \phi(t_n)$ for all $n \in \mathcal{N}_\tau$, and

$$\begin{aligned}\boldsymbol{\psi}(t) &:= \partial_t \boldsymbol{u}(t) - \frac{\boldsymbol{u}(t) + \boldsymbol{u}(t-\tau)}{\tau} = \frac{1}{\tau}\int_{t-\tau}^t (\xi - t + \tau)\partial_{\xi\xi}\boldsymbol{u}\,\mathrm{d}\xi,\\ \phi(t) &:= -\frac{1}{\lambda}\int_{t-\tau}^t \partial_\xi p\,\mathrm{d}\xi.\end{aligned}$$

Proceeding as in Step (i) of Exercise 75.2 shows that

$$\|\boldsymbol{\psi}_\tau\|_{\ell^2(J;\boldsymbol{L}^2)} \le \tau\|\partial_{tt}\boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}.$$

Moreover, it is clear that $\|\phi_\tau\|_{\ell^\infty(J;L^2)} \le \frac{\tau}{\lambda}\|\partial_t p\|_{C^0(\overline{J};L^2)}$. We also have

$$\partial_t\phi = -\frac{1}{\lambda}(\partial_t p(t) - \partial_t p(t-\tau)) = -\frac{1}{\lambda}\int_{t-\tau}^t \partial_{\xi\xi} p\,\mathrm{d}\xi,$$

and proceeding again as in Step (i) of Exercise 75.2, this implies that

$$\|\partial_t\phi\|_{L^2(J_*;L^2)} \le \lambda^{-1}\tau\|\partial_{tt} p\|_{L^2(J;L^2)}.$$

(ii.a) Since $\boldsymbol{e}^0 = \boldsymbol{0}$ and $r^0 = 0$, the stability estimate (75.9) implies that

$$\begin{aligned}\|\boldsymbol{e}_\tau\|^2_{\ell^\infty(J;\boldsymbol{L}^2)} + \mu\|\boldsymbol{e}_\tau\|^2_{\ell^2(J;\boldsymbol{V})} &\le c\,e^{\frac{4T}{\rho}}\tau^2\big(\rho\|\partial_{tt}\boldsymbol{u}\|^2_{L^2(J;\boldsymbol{L}^2)}\\ &\quad + (T+\rho)\mu^{-1}\|\partial_t p\|^2_{C^0(\overline{J};L^2)} + \mu^{-1}\rho^2\|\partial_{tt} p\|^2_{L^2(J;L^2)}\big),\end{aligned}$$

where we used that $\lambda := \lambda_0\mu$.
(ii.b) Let us now estimate the error on the pressure by invoking the stability estimate (75.10). We first observe that

$$\partial_t\boldsymbol{\psi} = \frac{1}{\tau}\int_{t-\tau}^t (\xi - t + \tau)\partial_{\xi\xi\xi}\boldsymbol{u}\,\mathrm{d}\xi, \quad \partial_t\phi = -\frac{1}{\lambda}\int_{t-\tau}^t \partial_{\xi\xi} p\,\mathrm{d}\xi.$$

This implies that

$$\|\partial_t \boldsymbol{\psi}\|_{L^2(J_*;\boldsymbol{L}^2)} \le \tau \|\partial_{ttt}\boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}, \qquad \|\partial_t \phi\|_{L^2(J_*;L^2)} \le \frac{\tau}{\lambda}\|\partial_{tt}p\|_{L^2(J;L^2)},$$
$$\|\delta_\tau \phi_\tau\|_{\ell^\infty(J_*;L^2)} \le \frac{\tau}{\lambda}\|\partial_{tt}p\|_{C^0(\overline{J_*};L^2)}, \quad \|\partial_{tt}\phi\|_{L^2(J_{**};L^2)} \le \frac{\tau}{\lambda}\|\partial_{ttt}p\|_{L^2(J_*;L^2)}.$$

Inserting these estimates in the stability estimate (75.10) implies that

$$\begin{aligned}\frac{1}{\mu}\|r_\tau\|^2_{\ell^2(J;L^2)} \le c\, e^{\frac{4T}{\rho}}\Big(&\rho^2\|\delta_\tau \boldsymbol{e}^1\|^2_{\boldsymbol{L}^2} + \frac{\tau}{\mu}\rho^2\|\delta_\tau r^1\|^2_{L^2}\\ &+ \rho\tau^2\big(\|\partial_{tt}\boldsymbol{u}\|^2_{L^2(J;\boldsymbol{L}^2)} + \rho^2\|\partial_{ttt}\boldsymbol{u}\|^2_{L^2(J;\boldsymbol{L}^2)}\big)\\ &+ \frac{\tau^2}{\mu}(T+\rho)\|\partial_t p\|^2_{C^0(\overline{J};L^2)} + \rho^2\|\partial_{tt}p\|^2_{C^0(\overline{J_*};L^2)}\big)\\ &+ \frac{\tau^2}{\mu}\rho^2\big(\|\partial_{tt}p\|^2_{L^2(J;L^2)} + \rho^2\|\partial_{ttt}p\|^2_{L^2(J_*;L^2)}\big)\Big).\end{aligned}$$

We need to estimate $\|\delta_\tau \boldsymbol{e}^1\|_{\boldsymbol{L}^2}$ and $\|\delta_\tau r^1\|_{L^2}$ to conclude. Recall that $\delta_\tau \boldsymbol{e}^1 = \frac{\boldsymbol{e}^1}{\tau}$ and $\delta_\tau r^1 = \frac{r^1}{\tau}$ since $\boldsymbol{e}^0 = \boldsymbol{0}$ and $r^1 = 0$. Using that $\boldsymbol{e}^0 = \boldsymbol{0}$ and $r^0 = 0$, we have

$$\frac{\boldsymbol{e}^1}{\tau} - \nabla\cdot(\mathbb{s}(\boldsymbol{e}^1)) + \nabla r^1 = \boldsymbol{\psi}^1, \quad \boldsymbol{e}^1_{|\partial D} = \boldsymbol{0}, \quad \frac{r^1}{\lambda} + \nabla\cdot\boldsymbol{e}^1 = \phi^1.$$

Hence, we have

$$\begin{aligned}&\frac{1}{2}\|\boldsymbol{e}^1\|^2_{\boldsymbol{L}^2} + 2\mu\tau\|\boldsymbol{e}^1\|^2_{\boldsymbol{V}} - \tau(\nabla\cdot\boldsymbol{e}^1, r^1)_{L^2} \le \frac{\tau^2}{2}\|\boldsymbol{\psi}^1\|^2_{\boldsymbol{L}^2},\\ &\frac{\tau}{2\lambda}\|r^1\|^2_{L^2} + \tau(\nabla\cdot\boldsymbol{e}^1, r^1)_{L^2} \le \frac{\tau\lambda}{2}\|\phi^1\|^2_{L^2},\end{aligned}$$

which gives

$$\frac{1}{2}\|\boldsymbol{e}^1\|^2_L + 2\mu\tau\|\boldsymbol{e}^1\|^2_{\boldsymbol{V}} + \frac{\tau}{2\lambda}\|r^1\|^2_{L^2} \le \frac{1}{2}\tau^3\Big(\|\partial_{tt}\boldsymbol{u}\|^2_{L^2(J_1;\boldsymbol{L}^2)} + \frac{1}{\lambda}\|\partial_t p\|^2_{C^0(\overline{J}_1;L^2)}\Big).$$

It is a this point that optimality is lost since this estimate implies that $\tau\|\delta_\tau \boldsymbol{e}^1\|^2_{\boldsymbol{L}} + \frac{\tau}{\lambda}\|\delta_t r^1\|^2_{L^2} \le c\tau(\|\partial_{tt}\boldsymbol{u}\|^2_{L^2(J_1;\boldsymbol{L}^2)} + \frac{1}{\lambda}\|\partial_t p\|^2_{C^0(\overline{J}_1;L^2)})$, i.e., the decay is $\mathcal{O}(\tau)$ instead of $\mathcal{O}(\tau^2)$. In conclusion, we obtain $\|r_\tau\|_{\ell^2(J;L^2)} \le c'(\boldsymbol{u}, p, T)\tau^{\frac{1}{2}}$.

Exercise 75.4 (Initialization). Using the hint, we denote by $(\boldsymbol{u}^{l,\gamma}, p^{l,\gamma})$ the approximation of the pair $(\boldsymbol{u}(t_{l+\gamma}), p(t_{l+\gamma}))$ using the first-order artificial compressibility algorithm (75.6) with the time step $\gamma\tau$ with $\gamma \in \{\frac{1}{3}, \frac{1}{2}, 1\}$ and $l \in \{1, 2\}$. Using Richardson's extrapolation technique, we have

$$\begin{aligned}\boldsymbol{u}^{l,\frac{1}{3}} &= \boldsymbol{u}(t_l) + c_1\frac{\tau}{3} + c_2\frac{\tau^2}{18} + \mathcal{O}(\tau^3),\\ \boldsymbol{u}^{l,\frac{1}{2}} &= \boldsymbol{u}(t_l) + c_1\frac{\tau}{2} + c_2\frac{\tau^2}{8} + \mathcal{O}(\tau^3),\\ \boldsymbol{u}^{l,1} &= \boldsymbol{u}(t_l) + c_1\tau + c_2\frac{\tau^2}{2} + \mathcal{O}(\tau^3).\end{aligned}$$

Similar expressions hold for the pressure. This gives

$$\begin{aligned}\boldsymbol{u}(t_l) &= \frac{9}{2}\boldsymbol{u}^{l,\frac{1}{3}} - 4\boldsymbol{u}^{l,\frac{1}{2}} + \frac{1}{2}\boldsymbol{u}^{l,1} + \mathcal{O}(\tau^3),\\ p(t_l) &= \frac{9}{2}p^{l,\frac{1}{3}} - 4p^{l,\frac{1}{2}} + \frac{1}{2}p^{l,1} + \mathcal{O}(\tau^3).\end{aligned}$$

Let us denote for all $l \in \{1,2\}$,

$$\begin{aligned}\boldsymbol{u}^{(l)} &:= \frac{9}{2}\boldsymbol{u}^{l,\frac{1}{3}} - 4\boldsymbol{u}^{l,\frac{1}{2}} + \frac{1}{2}\boldsymbol{u}^{l,1},\\ p^{(l)} &:= \frac{9}{2}p^{l,\frac{1}{3}} - 4p^{l,\frac{1}{2}} + \frac{1}{2}p^{l,1}.\end{aligned}$$

We obtain the approximations with the expected order as follows:

$$\begin{aligned}\partial_{tt}\boldsymbol{u}(\tau) &= \frac{\boldsymbol{u}^{(2)} - 2\boldsymbol{u}^{(1)} + \boldsymbol{u}_0}{\tau^2} + \mathcal{O}(\tau),\\ \partial_t\boldsymbol{u}(\tau) &= \frac{\boldsymbol{u}^{(2)} - \boldsymbol{u}_0}{\tau} + \mathcal{O}(\tau^2),\\ \boldsymbol{u}(\tau) &= \boldsymbol{u}^{(1)} + \mathcal{O}(\tau^3),\\ \partial_{tt}p(\tau) &= \frac{p^{(2)} - 2p^{(1)} + p(0)}{\tau^2} + \mathcal{O}(\tau),\\ \partial_t p(\tau) &= \frac{p^{(2)} - p(0)}{\tau} + \mathcal{O}(\tau^2),\\ p(\tau) &= p^{(1)} + \mathcal{O}(\tau^3).\end{aligned}$$

Chapter 76

Well-posedness and space semi-discretization

Exercises

Exercise 76.1 (Maximality). Let $V \hookrightarrow L$ be two real Hilbert spaces with norms $\|\cdot\|_V$ and $\|\cdot\|_L$. Let $R \in \mathcal{L}(V;L)$. Assume that R is a monotone operator, i.e., $\Re((R(v),v)_L) \geq 0$ for all $v \in V$. (i) Show that if R is maximal monotone (i.e., there is $\tau_0 > 0$ s.t. $I_V + \tau_0 R$ is surjective), then there are real numbers $c_1 > 0$ and $c_2 > 0$ s.t. $\sup_{w\in L} \frac{|(R(v),w)_L|}{\|w\|_L} \geq c_1\|v\|_V - c_2\|v\|_L$ for all $v \in V$. (*Hint*: show that $I_V + \tau_0 R$ is injective with closed image.) (ii) Show that if there are real numbers $c_1 > 0$ and $c_2 > 0$ s.t. $\sup_{w\in L} \frac{|(R(v),w)_L|}{\|w\|_L} \geq c_1\|v\|_V - c_2\|v\|_L$ for all $v \in V$, and $c_2 I_L + R^* : L' \equiv L \to V'$ is injective, then R is maximal monotone. (*Hint*: consider $S(v) := \sup_{w\in L} \frac{|(R(v)+c_2 v,w)_L|}{\|w\|_L}$ for all $v \in V$.) (iii) Assume that $I_V + \tau_0 R$ is surjective. Show that the norms $\|v\|_L + \tau_0\|R(v)\|_L$ and $\|v\|_V$ are equivalent.

Exercise 76.2 (Lemma 76.8). Revisit the proof of Lemma 76.8 by using Young's inequality in the form $a(s)\phi(s)^{\frac{1}{2}} \leq \frac{\theta a(s)^2}{4} + \frac{\phi(s)}{\theta}$, where θ is any time scale, and show that the choice $\theta = T$ leads to the sharpest estimate at the final time $t = T$. (*Hint*: minimize the function $\theta \mapsto \theta e^{\frac{T}{\theta}}$ at fixed T.)

Exercise 76.3 (Growth and decay in time). Assume that the linear operator $-\mu_\flat I_L + A \in \mathcal{L}(V_0;L)$ is maximal monotone where $\mu_\flat \in \mathbb{R}$, $\mu_\flat \neq 0$, but there is no constraint on the sign of $\mu_\flat$. Let $f \in C^0(\mathbb{R}_+;L)$ $\mathbb{R}_+ := [0,\infty)$. (i) Explain why there exists a unique $u \in C^1(\mathbb{R}_+;V_0)\cap C^0(\mathbb{R}_+;V_0)$ solving the problem $\partial_t u + A(u) = f$, $u(0) = u_0$. (ii) Assume now that $\mu_\flat > 0$. Show that the solution to this problem satisfies the following estimate for all $t \geq 0$:

$$\|u(t)\|_L^2 \leq e^{-\mu_\flat t}\|u_0\|_L^2 + \frac{1}{\mu_\flat}\int_0^t e^{-\mu_\flat(t-s)}\|f(s)\|_L^2 \,\mathrm{d}s.$$

(iii) Assume that $\mu_\flat > 0$ and $f \in C^0(\mathbb{R}_+;L) \cap L^\infty((0,\infty);L)$. Show that $\limsup_{t\to\infty} \|u(t)\|_L \leq \mu_\flat^{-1}\|f\|_{L^\infty((0,\infty);L)}$.

Exercise 76.4 (Wave equation). Consider the wave equation $\partial_{tt}p - \Delta p = g$ in $D\times J$ with the initial conditions $p(0) = p_0$ and $\partial_t p(0) = v_0$ in D and homogeneous Dirichlet conditions on p at the

boundary. Assume that $g \in L^2(D)$, $p_0, v_0 \in H^1_0(D)$, and $\Delta p_0 \in L^2(D)$. Show that this problem fits the setting of the time-dependent Friedrichs' systems from §76.3. (*Hint*: introduce $v := \partial_t p$ and $\boldsymbol{q} := -\nabla p$.)

Solution to exercises

Exercise 76.1 (Maximality). (i) Since R is maximal monotone, for every $f \in L$, there exists $u(f) \in V$ such that $u(f) + \tau_0 R(u(f)) = f$. But this $u(f)$ is unique since the monotonicity of R implies that $\|u(f)\|_L \le \|f\|_L$. Hence, the operator $I_V + \tau_0 R$ is bijective. In particular, this operator is injective and its image is closed, so that owing to Lemma C.39, we infer that there is $\alpha > 0$ s.t. $\sup_{w\in L} \frac{|(v+\tau_0 R(v),w)_L|}{\|w\|_L} \ge \alpha\|v\|_V$ for all $v \in V$. This also means that for all $v \in V$,

$$\begin{aligned}\sup_{w\in L} \frac{|(\tau_0 R(v),w)_L|}{\|w\|_L} &\ge \sup_{w\in L} \frac{|(v+\tau_0 R(v),w)_L|}{\|w\|_L} - \sup_{w\in L} \frac{|(v,w)_L|}{\|w\|_L}\\ &\ge \alpha\|v\|_V - \|v\|_L.\end{aligned}$$

This shows that $\sup_{w\in L} \frac{|(R(v),w)_L|}{\|w\|_L} \ge \alpha\tau_0^{-1}\|v\|_V - \tau_0^{-1}\|v\|_L$ for all $v \in V$.
(ii) We now prove the converse. Let us assume that there are real numbers $c_1 > 0$ and $c_2 > 0$ such that $\sup_{w\in L} \frac{|(R(v),w)_L|}{\|w\|_L} \ge c_1\|v\|_V - c_2\|v\|_L$ for all $v \in V$. Let us set $S(v) := \sup_{w\in L} \frac{|(c_2 v+R(v),w)_L|}{\|w\|_L}$ for all $v \in V$. Since R is monotone, we have

$$\begin{aligned}S(v) \ge \frac{|(c_2 v + R(v), v)_L|}{\|v\|_L} &= \frac{|c_2\|v\|_L^2 + (R(v),v)_L|}{\|v\|_L}\\ &\ge \frac{c_2\|v\|_L^2 + \Re((R(v),v)_L)}{\|v\|_L} \ge c_2\|v\|_L.\end{aligned}$$

Moreover, we have

$$S(v) \ge \sup_{w\in L} \frac{|(R(v),w)_L|}{\|w\|_L} - c_2\|v\|_L \ge c_1\|v\|_V - 2c_2\|v\|_L.$$

Hence, $3S(v) \ge c_1\|v\|_V$. This shows that the operator $T := c_2 I_V + R : V \to L$ is injective with closed image (here, we use $c_1 > 0$). Since T^* is injective (by assumption), this argument shows that the operator $c_2 I_V + R : V \to L$ is bijective. In particular, $c_2 I_V + R : V \to L$ is surjective, and so is the operator $I_V + c_2^{-1} R$ (here, we use $c_2 > 0$). We have shown that the operator is maximal monotone.
(iii) Using Step (i), we know that there is $\alpha > 0$ s.t.

$$\tau_0\|R(v)\|_L \ge \alpha\|v\|_V - \|v\|_L, \qquad \forall v \in V.$$

Hence, $\alpha\|v\|_V \le \|v\|_L + \tau_0\|R(v)\|_L$. Denoting $\iota_{L,V} := \sup_{v\in V} \frac{\|v\|_L}{\|v\|_V}$ (this number is finite since we assumed that V embeds continuously in L), we also have $\|v\|_L + \tau_0\|R(v)\|_L \le (\iota_{L,V} + \tau_0\|R\|_{\mathcal{L}(V;L)})\|v\|_V$.

Exercise 76.2 (Lemma 76.8). Young's inequality gives $a(s)\phi(s)^{\frac12} \le \frac{\theta a(s)^2}{4} + \frac{\phi(s)}{\theta}$. Hence, we obtain

$$\phi(t) \le \frac{\theta}{4}\|a\|^2_{L^2(0,t)} + b(t) + \int_0^t \frac{\phi(s)}{\theta}\,\mathrm{d}s.$$

Invoking Gronwall's lemma (see (65.2) from Exercise 65.3) with $\alpha(t) := \frac{\theta}{4}\|a\|^2_{L^2(0,t)} + b(t)$ and $\beta(t) := \frac{1}{\theta}$ shows that for all $t \in \overline{J}$,

$$\phi(t) \le e^{\frac{t}{\theta}}\big(\tfrac{\theta}{4}\|a\|^2_{L^2(0,t)} + b(t)\big).$$

In particular, at the final time $t = T$, we obtain

$$\phi(T) \le e^{\frac{T}{\theta}}\big(\tfrac{\theta}{4}\|a\|^2_{L^2(J)} + b(T)\big).$$

The sharpest bound is obtained by minimizing the function $\theta \mapsto \theta e^{\frac{T}{\theta}}$ (at fixed T), and computing the derivative shows that this function reaches its minimal value at $\theta = T$.

Exercise 76.3 (Growth and decay in time). (i) The Hille–Yosida theorem applied on the time interval $(0, T)$, where T is arbitrary, implies that there exists $v \in C^1(\mathbb{R}_+; V_0) \cap C^0(\mathbb{R}_+; V_0)$ so that

$$\partial_t v - \mu_\flat v + A(v) = e^{\mu_\flat t} f, \qquad v(0) = u_0.$$

Hence, we have

$$\partial_t(e^{-\mu_\flat t} v) + A(e^{-\mu_\flat t} v) = f, \qquad e^{-\mu_\flat \times 0} v(0) = 0.$$

Setting $u(t) := e^{-\mu_\flat t} v(t) \in C^1(\mathbb{R}_+; V_0) \cap C^0(\mathbb{R}_+; V_0)$ gives the unique solution to

$$\partial_t u + A(u) = f, \qquad u(0) = u_0.$$

(ii) Let $t \in \mathbb{R}_+$. Using u to test the equation $\partial_t u + A(u) = f$, which we recall holds true in $C^0(\mathbb{R}_+; L)$, using that $\mu_\flat > 0$ and $\Re\big(A(u), u\big)_L\big) \ge \mu_\flat \|u\|^2_L$, we infer that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|^2_L + \mu_\flat \|u(t)\|^2_L \le \|f(t)\|_L \|u(t)\|_L \le \frac{1}{2\mu_\flat}\|f(t)\|^2_L + \frac{\mu_\flat}{2}\|u(t)\|^2_L.$$

Hence, $\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\|^2_L + \mu_\flat \|u(t)\|^2_L \le \frac{1}{\mu_\flat}\|f(t)\|^2_L$. We obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(e^{\mu_\flat t}\|u(t)\|^2_L) \le \frac{1}{\mu_\flat} e^{\mu_\flat t}\|f(t)\|^2_L.$$

Recall that this inequality holds true in $C^0(\mathbb{R}_+; \mathbb{R})$. Integrating it over $(0, t)$, we infer that

$$\|u(t)\|^2_L \le e^{-\mu_\flat t}\|u_0\|^2_L + \frac{1}{\mu_\flat}\int_0^t e^{\mu_\flat(s-t)}\|f(s)\|^2_L \,\mathrm{d}s.$$

(iii) We still assume that $\mu_\flat > 0$. Since $f \in L^\infty((0,\infty); L)$, taking the square root on both sides in the above inequality and recalling that $J_t := (0, t)$, we infer that

$$\begin{aligned}
\|u(t)\|_L &\le e^{-\frac{\mu_\flat}{2}t}\|u_0\|_L + \mu_\flat^{-\frac{1}{2}}\|e^{\frac{1}{2}\mu_\flat(\cdot - t)} f\|_{L^2(J_t;L)} \\
&\le e^{-\frac{\mu_\flat}{2}t}\|u_0\|_L + \mu_\flat^{-\frac{1}{2}}\|e^{\mu_\flat(\cdot - t)}\|^{\frac{1}{2}}_{L^1(J_t)}\|f\|_{L^\infty(J_t;L)} \\
&\le e^{-\frac{\mu_\flat}{2}t}\|u_0\|_L + \frac{1}{\mu_\flat}\|f\|_{L^\infty(J_t;L)},
\end{aligned}$$

since $\|e^{\mu_\flat(\cdot - t)}\|_{L^1(J_t)} = \int_0^t e^{\mu_\flat(s-t)}\,\mathrm{d}s = \frac{1}{\mu_\flat}(1 - e^{-\mu_\flat t}) \le \frac{1}{\mu_\flat}$. The conclusion is straightforward since $\lim_{t\to\infty} e^{-\frac{\mu_\flat}{2}t} = 0$.

Exercise 76.4 (Wave equation). Following the hint and setting $u := (v, \boldsymbol{q})^\mathsf{T}$, we obtain $\partial_t v + \nabla{\cdot}\boldsymbol{q} = g$ and $\partial_t \boldsymbol{q} + \nabla v = \boldsymbol{0}$. Thus, we can rewrite the wave equation as $\partial_t u + A(u(t)) = f$ with $f := (g, \boldsymbol{0}) \in L := L^2(D;\mathbb{R}^m)$, $m := d+1$, and $A(u) := \sum_{k\in\{1:d\}} \mathcal{A}^k \partial_k u$ with

$$\mathcal{A}^k := \left[\begin{array}{c:c} 0 & \boldsymbol{e}_k \\ \hdashline \boldsymbol{e}_K^\mathsf{T} & \mathbb{O}_{d,d} \end{array}\right], \qquad \forall k \in \{1{:}d\},$$

where $(\boldsymbol{e}_k)_{k\in\{1:d\}}$ is the canonical basis of $\mathbb{R}^d$. Notice that $\mathcal{X} = \mathbb{O}_{m,m}$. The graph space is $V = H^1(D){\times}\boldsymbol{H}(\mathrm{div};D)$ and since we are enforcing a homogeneous Dirichlet condition on p and the initial condition p_0 is in $H^1_0(D)$, we have $v \in H^1_0(D)$. Hence, the solution u is sought in the space $C^1(\overline{J};L) \cap C^0(\overline{J};V_0)$ with $V_0 := H^1_0(D){\times}\boldsymbol{H}(\mathrm{div};D)$. Notice that $u_0 = (v_0, -\nabla p_0)^\mathsf{T} \in V_0$.

Chapter 77

Implicit time discretization

Exercises

Exercise 77.1 (Implicit advection-diffusion). Consider the 1D equation $\mu\partial_t u+\beta\partial_x u-\nu\partial_{xx}u = f$ in $D := (0,1)$, $t>0$, where $\mu\in\mathbb{R}_+$, $\beta\in\mathbb{R}$, $\nu\in\mathbb{R}_+$, $f\in L^2(D)$, boundary conditions $u(0)=0$, $u(1)=0$, and initial data $u_0=0$. Let $\mathcal{T}_h$ be the mesh composed of the cells $[ih,(i+1)h]$, $i\in\{0{:}I\}$, with uniform meshsize $h := \frac{1}{I+1}$. Let $V_h := P^{\mathrm{g}}_{1,0}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions that are zero at 0 and at 1 (see (19.37)). Let $(\varphi_i)_{i\in\{1:I\}}$ be the global Lagrange shape functions associated with the nodes $x_i := ih$ for all $i\in\{1{:}I\}$. (i) Write the fully discrete version of the problem in V_h using the implicit Euler time-stepping scheme. Denote the time step by τ and the discrete time nodes by $t_n := n\tau$ for all $n\in\mathcal{N}_\tau$. (ii) Prove a stability estimate. (*Hint*: consider the test function $2\tau u_h^n$ and introduce the Poincaré–Steklov constant C_{PS} s.t. $C_{\mathrm{PS}}\|v\|_{L^2(D)}\le \ell_D\|\partial_x v\|_{L^2(D)}$ for all $v\in H^1_0(D)$.) (iii) Letting $u_h^n := \sum_{i\in\{1:I\}}\mathsf{U}_i^n\varphi_i$ and $\mathsf{F}_i := \frac{1}{h}\int_D f\varphi_i\,\mathrm{d}x$ for all $i\in\{1{:}I\}$, write the linear system solved by the vector $\mathsf{U}^n := (\mathsf{U}_i^n)_{i\in\{1:I\}}$. (iv) Prove that $\max_{i\in\{1:I\}}\mathsf{U}_i^n \le \frac{\tau}{\mu}\max_{i\in\{1:I\}}\mathsf{F}_i + \max_{i\in\{1:I\}}\mathsf{U}_i^{n-1}$ if $\nu>|\beta|h$ and $\tau\ge\frac{\mu h^2}{3(2\nu-|\beta|h)}$. (*Hint*: consider the index $j\in\{1{:}I\}$ s.t. $\mathsf{U}_j^n=\max_{i\in\{1:I\}}\mathsf{U}_i^n$.)

Exercise 77.2 (Bound on $\|\dot e_h^1\|_L$). Prove (77.23). (*Hint*: use that $e_h^0=0$ and test (77.19) with $n:=1$ against $w_h := e_h^1$.)

Exercise 77.3 (IRK for advection-diffusion). Consider the advection-diffusion problem from Remark 77.7. Write the time-stepping process in functional and algebraic form using the IRK formalism from §69.2.4 and §70.1.3.

Exercise 77.4 (Implicit Euler, analysis using $\mathcal{P}_{V_h}$). The objective of this exercise is to derive an $\ell^\infty(\overline{J};L)$-error estimate for the implicit Euler scheme by using the operator $\mathcal{P}_{V_h}$ instead of the operator Π_h^{A} as was done in §77.3. We assume that $\tau\le\frac14\rho$. (i) Consider the following scheme: Given $u_h^0\in L$, one obtains $u_h^n\in V_h$ for all $n\in\mathcal{N}_\tau$ by solving

$$(u_h^n-u_h^{n-1},w_h)_L+\tau a_h(u_h^n,w_h)=\tau\phi^n(w_h),\qquad \forall w_h\in V_h,$$

with $\phi^n\in V_h'$. Set $\phi_\tau := (\phi^n)_{n\in\mathcal{N}_\tau}\in(V_h')^N$ and $\|\phi_\tau\|^2_{\ell^2((0,t_n);V_{h\flat}')} := \sum_{m\in\{1:n\}}\tau\|\phi^m\|^2_{V_{h\flat}'}$ with $\|\phi^m\|_{V_{h\flat}'} := \sup_{w_h\in V_h}\frac{|\phi^m(w_h)|}{\|w_h\|_{V_\flat}}$ and the norm $\|\cdot\|_{V_\flat}$ is defined as $\|v\|_{V_\flat} := \rho^{-1}\|v\|_L^2+\|v\|_{\mathcal{MS}}$ (this

is the definition used in the proof of Theorem 76.19; it differs from (77.16)). Show that for all $n \in \mathcal{N}_\tau$,

$$\|u_h^n\|_L \le e^{\frac{2t_n}{\rho}}\big(\|u_h^0\|_L + \|\phi_\tau\|_{\ell^2((0,t_n);V_{hb}')}\big).$$

(*Hint*: adapt the proof of Lemma 77.2.) (ii) Let $e_h^n := u_h^n - \mathcal{P}_{V_h}(u(t_n))$ and $\eta^n := \mathcal{P}_{V_h}(u(t_n)) - u(t_n)$ for all $n \in \overline{\mathcal{N}}_\tau$. Prove that $(e_h^n - e_h^{n-1}, w_h)_L + \tau a_h(e_h^n, w_h) = -\tau\phi^n(w_h)$ for all $w_h \in V_h$, with $\phi^n \in V_h'$ s.t.

$$\begin{aligned}\phi^n(w_h) &= (\psi^n + K(\eta^n) - \mathcal{X}\eta^n, w_h)_L + s_h(\mathcal{P}_{V_h}(u(t_n)), w_h)\\ &\quad + \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} + \mathcal{N})\eta^n, w_h)_{L(\partial D)} - (\eta^n, A_1(w_h))_L,\end{aligned}$$

and $\psi^n := \frac{1}{\tau}\int_{J_n}(\partial_t u(t) - \partial_t u(t_n))\mathrm{d}t \in L$. (*Hint*: see (76.27).) (iii) Let u solve (77.1) and let $u_{h\tau}$ solve (77.10). Assume that $u \in C^2(\overline{J}; L) \cap C^0(\overline{J}; H^{k+1}(D;\mathbb{C}^m))$. Prove that there is c s.t. for all $h \in \mathcal{H}$, all $\tau > 0$, and all $n \in \mathcal{N}_\tau$,

$$\|u(t_n) - u_h^n\|_L \le c\, e^{\frac{2t_n}{\rho}}\Big(\tau(\rho t_n)^{\frac12}c_1(t_n;u) + \big(h^{\frac12} + \big(\tfrac{t_n}{\rho}\big)^{\frac12}\max(\rho\beta,h)^{\frac12}h^{k+\frac12}\big)c_2(t_n;u)\Big),$$

with $c_1(t_n;u) := \|\partial_{tt}u\|_{C^0([0,t_n];L)}$ and $c_2(t_n;u) := |u|_{C^0([0,t_n];H^{k+1}(D;\mathbb{C}^m))}$. (*Hint*: see the proof of Theorem 76.19 and use Step (i).)

Solution to exercises

Exercise 77.1 (Implicit advection-diffusion). (i) Let τ be the time step. The fully discrete version of the problem in V_h using the implicit Euler time-stepping scheme is as follows: set $u_h^0 := 0$, then for all $n \in \mathcal{N}_\tau$ find $u_h^n \in V_h$ such that

$$\int_D \left(\mu\frac{u_h^n - u_h^{n-1}}{\tau}\varphi_i + \beta(\partial_x u_h^n)\varphi_i + \nu(\partial_x u_h^n)\partial_x\varphi_i\right)\mathrm{d}x = \int_D f\varphi_i\,\mathrm{d}x,$$

for all $i \in \{1{:}I\}$.
(ii) To establish a stability estimate, let us test the equation using $2\tau u_h^n$. This yields

$$\int_D \left(2\mu(u_h^n - u_h^{n-1})u_h^n + 2\tau\beta\frac{1}{2}\partial_x(u_h^n)^2 + 2\tau\nu(\partial_x u_h^n)^2\right)\mathrm{d}x = 2\tau\int_D f u_h^n\,\mathrm{d}x.$$

The term involving β vanishes owing to the boundary conditions. Moreover, using the identity $2(a-b)a = a^2 + (a-b)^2 - b^2$ and the Cauchy–Schwarz inequality, this gives

$$\mu\|u_h^n\|_{L^2}^2 + \mu\|u_h^n - u_h^{n-1}\|_{L^2}^2 - \mu\|u_h^{n-1}\|_{L^2}^2 + 2\tau\nu\|\partial_x u_h^n\|_{L^2}^2 \le 2\tau\|f\|_{L^2}\|u_h^n\|_{L^2}.$$

Using the inequality $2ab \le \lambda a^2 + \lambda^{-1}b^2$ with $\lambda := \nu C_{\mathrm{PS}}^2\ell_D^{-2}$, and dropping the nonnegative term $\mu\|u_h^n - u_h^{n-1}\|_{L^2}^2$ on the left-hand side, we have

$$\mu\|u_h^n\|_{L^2}^2 + 2\tau\nu\|\partial_x u_h^n\|_{L^2}^2 \le \mu\|u_h^{n-1}\|_{L^2}^2 + \frac{\tau}{\nu C_{\mathrm{PS}}^2\ell_D^{-2}}\|f\|_{L^2}^2 + \tau\nu C_{\mathrm{PS}}^2\ell_D^{-2}\|u_h^n\|_{L^2}.$$

Invoking the Poincaré–Steklov constant (i.e., $C_{\mathrm{PS}}\|u_h^n\|_{L^2} \le \ell_D\|\partial_x u_h^n\|_{L^2}$) for the rightmost term, we infer that

$$\mu\|u_h^n\|_{L^2}^2 + \tau\nu\|\partial_x u_h^n\|_{L^2}^2 \le \mu\|u_h^{n-1}\|_{L^2}^2 + \frac{\tau}{\nu C_{\mathrm{PS}}^2\ell_D^{-2}}\|f\|_{L^2}^2.$$

Summing the above inequality over $n \in \mathcal{N}_\tau$ and since $N\tau = T$, we obtain the following stability estimate:

$$\mu\|u_h^N\|_{L^2}^2 + \nu \sum_{n\in\mathcal{N}_\tau} \tau\|\partial_x u_h^n\|_{L^2}^2 \le \mu\|u_h^0\|_{L^2}^2 + \frac{T}{\nu C_{\mathrm{PS}}^2 \ell_D^{-2}}\|f\|_{L^2}^2.$$

(iii) The discrete system takes the following form: For all $i \in \{1{:}I\}$,

$$\mu\frac{h}{6\tau}(\mathsf{U}_{i-1}^n + 4\mathsf{U}_i^n + \mathsf{U}_{i+1}^n) + \frac{\beta}{2}(\mathsf{U}_i^n - \mathsf{U}_{i-1}^n) + \frac{\beta}{2}(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n) \\ + \frac{\nu}{h}(\mathsf{U}_i^n - \mathsf{U}_{i-1}^n) + \frac{\nu}{h}(\mathsf{U}_i^n - \mathsf{U}_{i+1}^n) = h\mathsf{F}_i + \mu\frac{h}{6\tau}(\mathsf{U}_{i-1}^{n-1} + 4\mathsf{U}_i^{n-1} + \mathsf{U}_{i+1}^{n-1}).$$

This can be simplified as follows:

$$\mu\frac{h}{6\tau}(\mathsf{U}_{i-1}^n + 4\mathsf{U}_i^n + \mathsf{U}_{i+1}^n) + \frac{\beta}{2}(\mathsf{U}_{i+1}^n - \mathsf{U}_{i-1}^n) \\ + \frac{\nu}{h}(-\mathsf{U}_{i-1}^n + 2\mathsf{U}_i^n - \mathsf{U}_{i+1}^n) = h\mathsf{F}_i + \mu\frac{h}{6\tau}(\mathsf{U}_{i-1}^{n-1} + 4\mathsf{U}_i^{n-1} + \mathsf{U}_{i+1}^{n-1}).$$

(iv) The above equation implies that

$$\mu\frac{h}{\tau}\mathsf{U}_i^n + \mu\frac{h}{6\tau}(\mathsf{U}_{i-1}^n - 2\mathsf{U}_i^n + \mathsf{U}_{i+1}^n) + \frac{\beta}{2}(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n + \mathsf{U}_i^n - \mathsf{U}_{i-1}^n) \\ + \frac{\nu}{h}(-\mathsf{U}_{i-1}^n + 2\mathsf{U}_i^n - \mathsf{U}_{i+1}^n) \le h\mathsf{F}^{\max} + \mu\frac{h}{\tau}\mathsf{U}^{n-1,\max},$$

where $\mathsf{F}^{\max} := \max_{i\in\{1:I\}} \mathsf{F}_i$, $\mathsf{U}^{m,\max} := \max_{i\in\{1:I\}} \mathsf{U}_i^m$, $m \in \{n-1, n\}$. The above expression can be rearranged as follows:

$$\mu\frac{h}{\tau}\mathsf{U}_i^n + \Big(\frac{\nu}{h} + \frac{\beta}{2} - \frac{\mu h}{6\tau}\Big)(\mathsf{U}_i^n - \mathsf{U}_{i-1}^n) + \Big(\frac{\nu}{h} - \frac{\beta}{2} - \frac{\mu h}{6\tau}\Big)(\mathsf{U}_i^n - \mathsf{U}_{i+1}^n) \le h\mathsf{F}^{\max} + \mu\frac{h}{\tau}\mathsf{U}^{n-1,\max}.$$

Let $j \in \{1{:}I\}$ be an index such that $\mathsf{U}_j^n = \mathsf{U}^{n,\max}$. Writing the above inequality for $i := j$, observing that $\mathsf{U}_j^n - \mathsf{U}_{j-1}^n \ge 0$, $\mathsf{U}_j^n - \mathsf{U}_{j+1}^n \ge 0$, and since $\frac{\nu}{h} + \frac{\beta}{2} - \frac{\mu h}{6\tau} \ge 0$ and $\frac{\nu}{h} - \frac{\beta}{2} - \frac{\mu h}{6\tau} \ge 0$ if $\nu > |\beta|h$ and $\tau \ge \frac{\mu h^2}{3(2\nu - |\beta|h)}$ (indeed, this last inequality is equivalent to $\frac{\nu}{h} \ge \frac{\mu h}{6\tau} + \frac{|\beta|}{2}$), we infer that

$$\mu\frac{h}{\tau}\mathsf{U}^{n,\max} = \mu\frac{h}{\tau}\mathsf{U}_j^n \le hF^{\max} + \mu\frac{h}{\tau}\mathsf{U}^{n-1,\max}.$$

This proves the assertion.

Exercise 77.2 (Bound on $\|\dot e_h^1\|_L$). Since $e_h^0 = 0$, we have $\dot e_h^1 = \frac{1}{\tau}e_h^1$. Testing (77.19) with $n := 1$ against $w_h := e_h^1$ yields

$$\|e_h^1\|_L^2 + \tau|e_h^1|_{\mathcal{MS}}^2 \le \tau\Lambda_\flat^-\|e_h^1\|_L^2 + \tau|(\alpha^1, e_h^1)_L|.$$

Since $\tau\Lambda_\flat^- \le \frac{\tau}{2\rho} \le \frac{1}{8}$, we infer that

$$\frac{7}{8}\|e_h^1\|_L^2 + \tau|e_h^1|_{\mathcal{MS}}^2 \le \tau\|\alpha^1\|_L\|e_h^1\|_L.$$

Hence, $\|e_h^1\|_L \le \frac{8}{7}\tau\|\alpha^1\|_L$, and it remains to estimate $\|\alpha^1\|_L$. We have

$$\begin{aligned}\|\alpha^1\|_L &\le \|\eta(\partial_t u)\|_{C^0(\overline{J}_1;L)} + \rho^{-1}\|\eta(u)\|_{C^0(\overline{J}_1;L)} + \tau\|\partial_{tt}u\|_{C^2(\overline{J}_1;L)}\\ &\le c\Big(\big(\tfrac{\beta}{\rho}\big)^{\frac12}h^{k+\frac12}c_2^1(u) + \tau c_1^1(u)\Big).\end{aligned}$$

Hence, we have

$$\|\dot e_h^1\|_L \le \tfrac{1}{\tau}\|e_h^1\|_L \le \tfrac{8}{7}\|\alpha^1\|_L \le c\Big(\tau c_1^1(u) + \big(\tfrac{\beta}{\rho}\big)^{\frac12}h^{k+\frac12}c_2^1(u)\Big).$$

The result is proved.

Exercise 77.3 (IRK for advection-diffusion). Let us consider an s-stage IRK scheme defined by its Butcher coefficients $\{a_{ij}\}_{i,j\in\{1:s\}}$, $\{b_i\}_{i\in\{1:s\}}$, $\{c_i\}_{i\in\{1:s\}}$, and its final stage coefficients $\{\alpha_i\}_{i\in\{0:s\}}$ defined in Remark 69.13. Let us set $t_{n,j} := t_{n-1} + c_j\tau$ for all $j \in \{1{:}s\}$ and all $n \in \mathcal{N}_\tau$. We then define $f_h(t) := \mathcal{P}_{V_h}(f(t))$ for all $t \in J$. We also define $A_h : X_h \to X_h$ by setting $(A_h(v_h), w_h)_L := \langle D(v_h), w_h\rangle_{X',X} + a_h(v_h, w_h)$ for all $v_h, w_h \in X_h$. The time stepping proceeds as follows. One first sets $u_h^0 := \mathcal{P}_{V_h}(u_0)$, then for all $n \in \mathcal{N}_\tau$ one seeks $\{u_h^{n,i}\}_{i\in\{1:s\}} \subset X_h$ solving the following system of coupled equations:

$$u_h^{n,i} - u_h^{n-1} = \tau\sum_{j\in\{1:s\}} a_{ij}\big(f_h(t_{n,j}) - A_h(u_h^{n,j})\big),$$

and the update at t_n is obtained by setting $u_h^n := \alpha_0 u_h^{n-1} + \sum_{p\in\{1:s\}}\alpha_p u_h^{n,p}$. Let $\{\varphi_i\}_{i\in\{1:I\}}$ be a basis of X_h. Recalling the mass matrix $\mathcal{M} \in \mathbb{C}^{I\times I}$ and the stiffness matrix $\mathcal{A} \in \mathbb{C}^{I\times I}$ introduced in §77.1.2, the algebraic realization of the IRK scheme proceeds as follows: One first lets $\mathsf{U}^0 \in \mathbb{C}^I$ be the coordinate vector of $\mathcal{P}_{V_h}(u_0)$. Then, for all $n \in \mathcal{N}_\tau$, letting $\mathsf{U}^{n,p} \in \mathbb{C}^I$ be the coordinate vector of $u_h^{n,p}$ for all $p \in \{1{:}s\}$, the IRK scheme consists of solving the linear system:

$$\begin{pmatrix}\mathcal{M}+\tau a_{11}\mathcal{A} & \cdots & \tau a_{1s}\mathcal{A}\\ \vdots & \ddots & \vdots\\ \tau a_{s1}\mathcal{A} & \cdots & \mathcal{M}+\tau a_{ss}\mathcal{A}\end{pmatrix}\begin{pmatrix}\mathsf{U}^{n,1}\\ \vdots\\ \mathsf{U}^{n,s}\end{pmatrix} = \begin{pmatrix}\mathsf{G}^{n,1}\\ \vdots\\ \mathsf{G}^{n,s}\end{pmatrix},$$

with the load vectors defined by $\mathsf{G}^{n,p} := \mathcal{M}\mathsf{U}^{n-1} + \tau\sum_{q\in\{1:s\}} a_{pq}\mathsf{F}^{n,q} \in \mathbb{C}^I$, and $\mathsf{F}_i^{n,p} := (f_h(t_{n,p}), \varphi_i)_L$ for all $p \in \{1{:}s\}$. Finally, one sets $\mathsf{U}^n := \alpha_0\mathsf{U}^{n-1} + \sum_{p\in\{1:s\}}\alpha_p\mathsf{U}^{n,p}$.

Exercise 77.4 (Implicit Euler, analysis using $\mathcal{P}_{V_h}$). Notice that $\tau \le \frac14\rho$ implies that $\tau\Lambda_\flat^- < \frac18 < 1$, so that the discrete problem is well-posed.
(i) We use $w_h := u_h^n$ as the test function, take the real part, invoke the lower bound (77.5) and the identity $(u_h^n - u_h^{n-1}, u_h^n)_L = \frac12\|u_h^n\|_L^2 - \frac12\|u_h^{n-1}\|_L^2 + \frac12\|u_h^n - u_h^{n-1}\|_L^2$. This gives

$$\frac12\|u_h^n\|_L^2 + \tau\Lambda_\flat\|u_h^n\|_L^2 + \tau|u_h^n|_{\mathcal{MS}}^2 \le \frac12\|u_h^{n-1}\|_L^2 + \tau|\phi^n(u_h^n)|.$$

Since $|\phi^n(u_h^n)| \le \|\phi^n\|_{V'_{h\flat}}\|u_h^n\|_{V_\flat} \le \frac12\|\phi^n\|_{V'_{h\flat}}^2 + \frac12\|u_h^n\|_{V_\flat}^2 = \frac12\|\phi^n\|_{V'_{h\flat}}^2 + \frac{1}{2\rho}\|u_h^n\|_L^2 + \frac12|u_h^n|_{\mathcal{MS}}^2$ and since $\frac{1}{2\rho} - \Lambda_\flat \le \frac{1}{\rho}$, this gives

$$\frac12\|u_h^n\|_L^2 + \frac12\tau|u_h^n|_{\mathcal{MS}}^2 \le \frac12\|u_h^{n-1}\|_L^2 + \frac{\tau}{\rho}\|u_h^n\|_L^2 + \frac12\tau\|\phi^n\|_{V'_{h\flat}}^2.$$

Dropping the nonnegative term $\frac{1}{2}\tau|u_h^n|_{\mathcal{MS}}^2$ from the left-hand side and summing the inequalities for all $m \in \{1:n\}$ gives

$$\|u_h^n\|_L^2 \le \|u_h^0\|_L^2 + \sum_{m\in\{1:n\}} \frac{2\tau}{\rho}\|u_h^m\|_L^2 + \sum_{m\in\{1:n\}} \tau\|\phi^m\|_{V_{h\flat}'}^2.$$

We apply the discrete Gronwall lemma from Exercise 68.3 with $\gamma := \frac{2\tau}{\rho} \in (0,1)$ by assumption, $a_m := \|u_h^m\|_L^2$, $b_m := 0$, $c_m := \tau\|\phi^m\|_{V_{h\flat}'}^2$, and $B := \|u_h^0\|_L^2$. Since $\gamma \le \frac{1}{2}$ by assumption, we have $\frac{1}{1-\gamma} \le e^{2\gamma}$. This completes the proof of the assertion.
(ii) Subtracting (77.1) from (77.10) gives

$$(e_h^n - e_h^{n-1}, w_h)_L + \tau\big(a_h(u_h^n, w_h) - (A(u(t_n), w_h)_L\big) = -\tau(\psi^n, w_h)_L,$$

for all $w_h \in V_h$. This implies that

$$(e_h^n - e_h^{n-1}, w_h)_L + \tau a_h(e_h^n, w_h) = -\tau\phi^n(w_h),$$

with

$$\phi^n(w_h) := (\psi^n, w_h)_L + \big(a_h(\mathcal{P}_{V_h}(u(t_n)), w_h) - (A(u(t_n)), w_h)_L\big).$$

Proceeding as in the derivation of (76.27), we can rearrange the second term on the right-hand side to obtain the expected expression for $\phi^n(w_h)$.
(iii) We now estimate $\|\phi^n\|_{V_{h\flat}'}$. Let us denote by $(\phi_i^n)_{i\in\{1:4\}}$ the four antilinear forms composing ϕ^n, i.e.,

$$\phi_1^n(w_h) := (\psi^n + K(\eta^n) - \mathcal{X}\eta^n, w_h)_L, \quad \phi_1^n(w_h) := s_h(\mathcal{P}_{V_h}(u(t_n)), w_h),$$
$$\phi_3^n(w_h) := \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} + \mathcal{N})\eta^n, w_h)_{L(\partial D)}, \quad \phi_4^n(w_h) := -(\eta^n, A_1(w_h))_L.$$

Since we assumed that $u \in C^2(\overline{J}; L)$, we have $\|\psi^n\|_L \le \tau\|\partial_{tt}u\|_{C^0(\overline{J}_n;L)}$, and our simplifying assumption on K and $\mathcal{X}$ implies that $\|K(\eta^n) - \mathcal{X}\eta^n\|_L \le c\rho^{-1}\|\eta^n\|_L$. Since $\rho^{-\frac{1}{2}}\|\cdot\|_L \le \|\cdot\|_{V_\flat}$, invoking the Cauchy–Schwarz inequality gives

$$\begin{aligned}\|\phi_1^n\|_{V_{h\flat}'} &\le \rho^{\frac{1}{2}}(\tau\|\partial_{tt}u\|_{C^0(\overline{J}_n;L)} + \rho^{-1}\|\eta^n\|_L)\\ &\le \rho^{\frac{1}{2}}\tau\|\partial_{tt}u\|_{C^0(\overline{J}_n;L)} + c\,\rho^{-\frac{1}{2}}h^{k+1}|u(t_n)|_{H^{k+1}},\end{aligned}$$

where we used the approximation properties of $\mathcal{P}_{V_h}$ (see Propositions 22.19 and 22.21 and recall that we are assuming that the mesh sequence is quasi-uniform). The assumption (76.20b) on s_h implies that

$$\|\phi_2^n\|_{V_{h\flat}'} \le c\beta^{\frac{1}{2}}h^{k+\frac{1}{2}}|u(t_n)|_{H^{k+1}}.$$

The assumption (76.19c) on $\mathcal{M}^{\mathrm{BP}}$ and the approximation properties of $\mathcal{P}_{V_h}$ imply that

$$\|\phi_3^n\|_{V_{h\flat}'} \le c\beta^{\frac{1}{2}}h^{k+\frac{1}{2}}|u(t_n)|_{H^{k+1}}.$$

Finally, using that $|w_h|_{\mathcal{S}} + \big(\frac{\beta}{\ell_D}\big)^{\frac{1}{2}}\|w_h\|_L \le \|w_h\|_{V_\flat}$, the assumption (76.20c) on s_h yields

$$\|\phi_4^n\|_{V_{h\flat}'} \le c\beta^{\frac{1}{2}}h^{k+\frac{1}{2}}|u(t_n)|_{H^{k+1}}.$$

Putting these bounds together yields

$$\|\phi^n\|_{V_{h\flat}'} \le \rho^{\frac{1}{2}}\tau\|\partial_{tt}u\|_{C^0(\overline{J}_n;L)} + c\,\rho^{-\frac{1}{2}}\max(\rho\beta, h)^{\frac{1}{2}}h^{k+\frac{1}{2}}|u(t_n)|_{H^{k+1}}.$$

We can now apply Step (i) and since $e_h^0 = 0$, we infer that for all $n \in \mathcal{N}_\tau$,

$$\|e_h^n\|_L \le c\, e^{\frac{2t_n}{\rho}} \Big(\tau(\rho t_n)^{\frac12} \|\partial_{tt} u\|_{C^0([0,t_n];L)} + (\rho^{-1} t_n)^{\frac12} \max(\rho\beta, h)^{\frac12} h^{k+\frac12} |u|_{C^0([0,t_n];H^{k+1})} \Big),$$

where we used that $\sum_{m\in\{1:n\}} \tau |u(t_m)|_{H^{k+1}}^2 \le t_n |u|_{C^0([0,t_n];H^{k+1})}^2$. The conclusion follows by using the triangle inequality.

Chapter 78

Explicit time discretization

Exercises

Exercise 78.1 (Order conditions). (i) Consider the linear ODE system $\partial_t \mathsf{U} = \tilde{\mathcal{A}}\mathsf{U} + \tilde{\mathsf{F}}$. Let $p \geq 1$. Prove that

$$\mathsf{U}(t_n) = \sum_{r\in\{0:p\}} \frac{\tau^r}{r!}\tilde{\mathcal{A}}^r\mathsf{U}(t_{n-1}) + \tau\mathsf{G}_p(t_{n-1}) + \mathcal{O}(\tau^{p+1}), \tag{78.1}$$

with G_p defined in (78.13). (*Hint*: verify that $\partial_t^r \mathsf{U} = \tilde{\mathcal{A}}^r\mathsf{U} + \Phi_r(\tilde{\mathsf{F}})$ for all $r \geq 1$, with $\Phi_r(\tilde{\mathsf{F}}) := \sum_{q\in\{1:r\}} \tilde{\mathcal{A}}^{r-q}\partial_t^{q-1}\tilde{\mathsf{F}}$.) (ii) Let $\tilde{\mathsf{F}} \in C^\infty(\overline{J};\mathbb{C}^I)$. Consider the uncoupled ODE system $\partial_t\mathsf{U} = \tilde{\mathsf{F}}(t)$. Let $\mathsf{U}^{n-1} := \mathsf{U}(t_{n-1})$. Let U^n be given by the RK scheme. Show that a necessary and sufficient condition for $\mathsf{U}(t_n) - \mathsf{U}^n = \mathcal{O}(\tau^{p+1})$ is (78.10) with $r := 1$. (*Hint*: write a Taylor expansion of order $(p-1)$ of $\tilde{\mathsf{F}}(t_{n,j})$ for all $j \in \{1:s\}$.)

Exercise 78.2 (Condition (78.10)). (i) Show that if (78.9a) holds true, then $\sum_{j\in\{1:s\}} b_j(1-c_j)^m c_j^n = \frac{m!n!}{(m+n+1)!}$ for all $m,n \in \mathbb{N}$ s.t. $m+n \leq p-1$. (*Hint*: recall that $(1+x)^m = \sum_{r\in\{0:m\}}\binom{m}{r}x^r$, $\frac{1}{n+l+1} = \int_0^1 x^{n+l}\,\mathrm{d}x$, and $\int_0^1(1-x)^m x^n\,\mathrm{d}x = \frac{m!n!}{(m+n+1)!}$.) (ii) Show that if (78.9a) and (78.9c) hold true, then $\sum_{i\in\{1:s\}} b_i(1-c_i)^{m-1}a_{ij} = \frac{b_j}{m}(1-c_j)^m$ for all $j \in \{1:s\}$ and all $m \in \{1:\zeta\}$. (iii) Prove that (78.10) is met for $q := 1$ if (78.9a) and (78.9b) hold with $\eta := p-1$. (*Hint*: show that $\sum_{j_2,\dots,j_r\in\{1:s\}} a_{j_1j_2}\dots a_{j_{r-1}j_r} = \frac{1}{(r-1)!}c_{j_1}^{r-1}$ for all $r \in \{2:p\}$.) (iv) Prove that (78.10) is met for $q := 1$ if (78.9a) and (78.9c) hold with $\zeta := p-1$. (v) Show that (78.10) with $q := 1$ is met for all $r \in \{1:p\}$ if (78.9a) holds and (78.9b) and (78.9c) hold with $\eta+\zeta+1 = p$ (vi) Show that (78.10) is met for all $r \in \{1:p\}$ and all $q \in \{1:p-r+1\}$ if (78.9a) holds and (78.9b) and (78.9c) hold with $p \leq \eta+\zeta+1$.

Exercise 78.3 (Explicit Euler). Revisit the proof of Lemma 78.12 by using the test function $w_h := u_h^n$ instead of $w_h := u_h^{n-1}$ and assuming that $\tau \leq \min(\lambda_0\tau_2(h), \frac{1}{2}\frac{\rho}{1+\lambda_0\varpi^2})$ where $\varpi := \frac{h}{\beta}\sup_{v_h,w_h\in V_h}\frac{|a_h(v_h,w_h)|}{\|v_h\|_L\|w_h\|_L}$. (*Hint*: use that $a_h(u_h^{n-1},u_h^n) = a_h(u_h^n,u_h^n) + a_h(u_h^{n-1}-u_h^n,u_h^n)$.)

Exercise 78.4 (First-order viscosity). Let $(\cdot,\cdot)_V$ be a semidefinite Hermitian sesquilinear form in V and let $|\cdot|_V$ be the associated seminorm. Assume that $\Re\big((A(v),v)_L\big) \geq 0$ and $\|A(v)\|_L \leq \beta\|v\|_L$ for all $v \in V$. Let $V_h \subset V$ and set $c_{\mathrm{INV}}(h) := \max_{v_h\in V_h}\frac{|v_h|_V}{\|v_h\|_L}$. Given $u_h^0 \in V_h$, let $u_h^n \in V_h$

solve $\frac{1}{\tau}(u_h^n - u_h^{n-1}, w_h)_L + (A(u^{n-1}), v_h)_L + \mu(u_h^{n-1}, w_h)_V = 0$, for all $w_h \in V_h$ and all $n \in \mathcal{N}_\tau$, where $\mu \geq 0$ is an artificial viscosity parameter yet to be defined (μ can depend on h and τ). (i) Explain why this scheme can be more attractive than the implicit Euler method with $\mu := 0$. (ii) Prove that if $\tau(\beta + \mu c_{\text{INV}}(h))^2 \leq 2\mu$, then $\|u_h^n\|_L \leq \|u_h^0\|_L$ for all $n \in \mathcal{N}_\tau$. (iii) Prove that the above stability condition can be realized if and only if $2\beta\tau c_{\text{INV}}(h) \leq 1$, and determine the admissible range for μ. *Note*: the constant $\beta\tau c_{\text{INV}}(h)$ is called Courant–Friedrichs–Levy (CFL) number.

Exercise 78.5 (Explicit Euler, mass lumping). Let $\beta \in \mathbb{R}$, $\beta \neq 0$. Consider the equation $\partial_t u + \beta \partial_x u = 0$ over $D := (0,1)$ with periodic boundary conditions. Use the same setting for the space discretization as in Exercise 77.1. (i) Write the linear system solved by the coordinate vector $(\mathsf{U}_1^n, \dots, \mathsf{U}_I^n)^\mathsf{T}$ by using the explicit Euler scheme and the Galerkin approximation with mass lumping. (*Hint*: use the convention $\mathsf{U}_I^n := \mathsf{U}_0^n$, $\mathsf{U}_{I+1}^n := \mathsf{U}_1^n$, $\mathsf{U}_{-1}^n := \mathsf{U}_{I-1}^n$.) (ii) Show that $\sum_{j\in\{1:I\}}(\mathsf{U}_j^n)^2 = \sum_{j\in\{1:I\}}(\mathsf{U}_j^{n-1})^2 + \lambda^2 \sum_{j\in\{1:I\}}(\mathsf{U}_{j+1}^{n-1} - \mathsf{U}_{j-1}^{n-1})^2$ with $\lambda := \frac{\beta\tau}{2h}$. (iii) Let $a := (1 - 2\mathrm{i}\lambda \sin(\frac{k}{I}2\pi))$ where $k \in \mathbb{N}$ and $\frac{k}{I} \notin \mathbb{N}$, $\mathrm{i}^2 := -1$, and set $\mathsf{U}_j^0 := ae^{\mathrm{i}\frac{j}{I}2k\pi}$ for all $j \in \{1:I\}$. Compute U_j^n for all $n \in \mathcal{N}_\tau$ and comment on the result.

Exercise 78.6 (Error equation, RK2). (i) Verify that

$$u(t_n) = u(t_{n-1}) + \tau\partial_t u(t_{n-1}) + \frac{1}{2}\tau^2\partial_{tt}u(t_{n-1}) + \frac{1}{2}\tau\psi^{n-1},$$

with $\psi^{n-1} := \frac{1}{\tau}\int_{J_n}(t_n - t)^2\partial_{ttt}u(t)\mathrm{d}t$. (*Hint*: integrate by parts in time.) (ii) Prove (78.26). (*Hint*: use the fact that $(\partial_{tt}u(t_{n-1}), w_h)_L + (A(\partial_t u(t_{n-1})), w_h)_L = (\partial_t f^{n-1}, w_h)_L$ for all $w_h \in V_h$.)

Exercise 78.7 (ERK schemes, $p = 3$). Prove Lemma 78.21. (*Hint*: proceed as in the proof of Lemma 78.15, use that $\|A_h(w_h)\|_L \leq c\frac{1}{\rho}\|w_h\|_{H^1}$ for all $w_h \in V_h$, and invoke the H^1-stability of $\mathcal{P}_{V_h}$ (see Proposition 22.21).)

Solution to exercises

Exercise 78.1 (Order conditions). (i) We verify the hint by induction on $r \geq 1$. The assertion is satisfied for $r = 1$ since $\Phi_1(\tilde{\mathsf{F}}) = \tilde{\mathsf{F}}$, so that we indeed have $\partial_t\mathsf{U} = \tilde{\mathcal{A}}\mathsf{U} + \Phi_1(\tilde{\mathsf{F}})$. Assume that the assertion is satisfied for some $r \geq 1$, and let us show that it holds true for $(r+1)$. We have

$$\begin{aligned}
\partial_t^{r+1}\mathsf{U} &= \partial_t\big(\partial_t^r\mathsf{U}\big) = \partial_t\big(\tilde{\mathcal{A}}^r\mathsf{U} + \Phi_r(\tilde{\mathsf{F}})\big) \\
&= \tilde{\mathcal{A}}^r\partial_t\mathsf{U} + \sum_{q\in\{1:r\}}\tilde{\mathcal{A}}^{r-q}\partial_t^q\tilde{\mathsf{F}} \\
&= \tilde{\mathcal{A}}^r(\tilde{\mathcal{A}}\mathsf{U} + \tilde{\mathsf{F}}) + \sum_{q\in\{2:r+1\}}\tilde{\mathcal{A}}^{r+1-q}\partial_t^{q-1}\tilde{\mathsf{F}} \\
&= \tilde{\mathcal{A}}^{r+1}\mathsf{U} + \sum_{q\in\{1:r+1\}}\tilde{\mathcal{A}}^{r+1-q}\partial_t^{q-1}\tilde{\mathsf{F}}.
\end{aligned}$$

This proves the assertion on the time derivatives of U. The Taylor expansion of order p of U at t_n then becomes

$$\begin{aligned}\mathsf{U}(t_n) &= \mathsf{U}(t_{n-1}) + \sum_{r\in\{1:p\}} \frac{\tau^r}{r!}\partial_t^r \mathsf{U}(t_{n-1}) + \mathcal{O}(\tau^{p+1}) \\ &= \mathsf{U}(t_{n-1}) + \sum_{r\in\{1:p\}} \frac{\tau^r}{r!}\big(\tilde{\mathcal{A}}^r \mathsf{U}(t_{n-1}) + \Phi_r(\tilde{\mathsf{F}})(t_{n-1})\big) + \mathcal{O}(\tau^{p+1}) \\ &= \sum_{r\in\{0:p\}} \frac{\tau^r}{r!}\tilde{\mathcal{A}}^r \mathsf{U}(t_{n-1}) + \tau \mathsf{G}_p(t_{n-1}) + \mathcal{O}(\tau^{p+1}),\end{aligned}$$

since the definition (78.13) gives

$$\sum_{r\in\{1:p\}} \frac{\tau^r}{r!}\Phi_r(\tilde{\mathsf{F}})(t) = \sum_{r\in\{1:p\}} \frac{\tau^r}{r!} \sum_{q\in\{1:r\}} \tilde{\mathcal{A}}^{r-q}\partial_t^{q-1}\tilde{\mathsf{F}}(t) = \tau \mathsf{G}_p(t).$$

(ii) By definition, we have

$$\mathsf{U}^n = \mathsf{U}^{n-1} + \tau \sum_{j\in\{1:s\}} b_j \tilde{\mathsf{F}}(t_{n,j}).$$

But $\tilde{\mathsf{F}}(t_{n,j}) = \sum_{r\in\{0:p-1\}} \frac{(c_j\tau)^r}{r!}\partial_t^r \tilde{\mathsf{F}}(t_{n-1}) + \mathcal{O}(\tau^p)$. Recalling that $\partial_t^{r+1}\mathsf{U} = \partial_t^r \tilde{\mathsf{F}}$, this means that

$$\begin{aligned}\mathsf{U}^n &= \mathsf{U}^{n-1} + \tau \sum_{j\in\{1:s\}} b_j \sum_{q\in\{0:p-1\}} \frac{(c_j\tau)^q}{q!}\partial_t^q \tilde{\mathsf{F}}(t_{n-1}) + \mathcal{O}(\tau^{p+1}) \\ &= \mathsf{U}^{n-1} + \sum_{q\in\{0:p-1\}} \frac{\tau^{q+1}}{(q+1)!}\partial_t^{q+1}\mathsf{U}(t_{n-1}) \sum_{j\in\{1:s\}} (q+1) b_j c_j^r + \mathcal{O}(\tau^{p+1}) \\ &= \mathsf{U}^{n-1} + \sum_{q\in\{1:p\}} \frac{\tau^q}{q!}\partial_t^q \mathsf{U}(t_{n-1}) \sum_{j\in\{1:s\}} q b_j c_j^{q-1} + \mathcal{O}(\tau^{p+1}).\end{aligned}$$

Hence, we have $\mathsf{U}^n - \mathsf{U}(t_n) = \mathcal{O}(\tau^{p+1})$ iff the above identity coincides with the Taylor expansion of order p of $\mathsf{U}(t_n)$ at t_{n-1}. This is true iff $\sum_{j\in\{1:s\}} q b_j c_j^{q-1} = 1$ for all $q \in \{1{:}p\}$. This is exactly (78.10) with $r := 1$.

Exercise 78.2 (Condition (78.10)**).** (i) Using (78.9a) and the binomial formula, we obtain

$$\begin{aligned}\sum_{j\in\{1:s\}} b_j (1-c_j)^m c_j^n &= \sum_{l\in\{0:m\}} \binom{m}{l} (-1)^l \sum_{j\in\{1:s\}} b_j c_j^{l+n} \\ &= \sum_{l\in\{0:m\}} \binom{m}{l} (-1)^l \frac{1}{n+l+1},\end{aligned}$$

where we used that $n + l + 1 \le n + m + 1 \le p$ to invoke (78.9a). But

$$\begin{aligned}\sum_{l\in\{0:m\}} \binom{m}{l} (-1)^l \frac{1}{n+l+1} &= \sum_{l\in\{0:m\}} \binom{m}{l} (-1)^l \int_0^1 x^{n+l}\,\mathrm{d}x \\ &= \int_0^1 (1-x)^m x^n\,\mathrm{d}x = \frac{m!n!}{(m+n+1)!}.\end{aligned}$$

This shows that

$$\sum_{j\in\{1:s\}} b_j(1-c_j)^m c_j^n = \frac{m!n!}{(m+n+1)!}.$$

Notice in passing that this proves that the quadrature with the nodes $\{c_i\}_{i\in\{1:s\}}$ and weights $\{b_i\}_{i\in\{1:s\}}$ is at least of order $p-1$ since it integrates exactly the Bernstein basis $\{\binom{p-1}{m}(1-x)^m x^{p-1-m}\}_{m\in\{0:p-1\}}$.
(ii) To prove the second identity, we proceed as above and use (78.9c) to infer that for all $j\in\{1:s\}$ and all $m\in\{1:\zeta\}$,

$$\begin{aligned}
\sum_{i\in\{1:s\}} b_i(1-c_i)^{m-1}a_{ij} &= \sum_{r\in\{0:m-1\}}\binom{m-1}{r}\sum_{i\in\{1:s\}} b_i(-c_i)^r a_{ij}\\
&= \sum_{r\in\{0:m-1\}}\binom{m-1}{r}(-1)^r\frac{b_j}{r+1}(1-c_j^{r+1})\\
&= \frac{b_j}{m}\sum_{r\in\{0:m-1\}}\binom{m}{r+1}(-1)^r(1-c_j^{r+1})\\
&= -\frac{b_j}{m}\sum_{r\in\{1:m\}}\binom{m}{r}(-1)^r(1-c_j^r) = -\frac{b_j}{m}\sum_{r\in\{1:m\}}\binom{m}{r}\big((-1)^r-(-c_j)^r\big)\\
&= -\frac{b_j}{m}\sum_{r\in\{0:m\}}\binom{m}{r}\big((-1)^r-(-c_j)^r\big)\\
&= -\frac{b_j}{m}(0-(1-c_j)^m) = \frac{b_j}{m}(1-c_j)^m.
\end{aligned}$$

(iii) Let us now show that (78.10) with $q:=1$ is met for all $r\in\{1:p\}$ if (78.9a) holds and (78.9b) holds with $\eta:=p-1$. For $r:=1$, (78.10) boils down to $\sum_{j\in\{1:s\}} b_j = 1$, which is nothing but (78.9a) with $q:=1$. Let now $r\in\{2:p\}$. Using (78.9b) for all $q\in\{1:r-1\}$ (this is legitimate since $r-1\le p-1=\eta$), we obtain

$$\begin{aligned}
\sum_{j_2,\dots,j_r\in\{1:s\}} a_{j_1j_2}\times\ldots\times a_{j_{r-1}j_r} &= \sum_{j_2,\dots,j_{r-1}\in\{1:s\}} a_{j_1j_2}\times\ldots\times a_{j_{r-2}j_{r-1}}\sum_{j_r\in\{1:s\}} a_{j_{r-1}j_r}\\
&= \sum_{j_2,\dots,j_{r-1}\in\{1:s\}} a_{j_1j_2}\times\ldots\times a_{j_{r-2}j_{r-1}}c_{j_{r-1}}\\
&= \sum_{j_2,\dots,j_{r-2}\in\{1:s\}} a_{j_1j_2}\times\ldots\times a_{j_{r-3}j_{r-2}}\frac{1}{2}c_{j_{r-2}}^2\\
&= \ldots = \frac{1}{(r-1)!}c_{j_1}^{r-1}.
\end{aligned}$$

Now, invoking (78.9a) with $q:=r$ (this is legitimate since $r\le p$) gives

$$\sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1}a_{j_1j_2}\times\ldots\times a_{j_{r-1}j_r} = \frac{1}{r!}.$$

(iv) Let us now show that (78.10) with $q:=1$ is met for all $r\in\{1:p\}$ if (78.9a) holds and (78.9c) holds with $\zeta:=p-1$. For $r:=1$, (78.10) boils down to $\sum_{j\in\{1:s\}} b_j = 1$, which is nothing but

(78.9a) with $q := 1$. Let now $r \in \{2{:}p\}$. Using (78.9c) (actually the expression from Step (ii) for all $m \in \{1{:}r{-}1\}$; this is legitimate since $r - 1 \le p - 1 = \zeta$), we obtain

$$\begin{aligned}\sum_{j_1\in\{1:s\}} \sum_{j_2,\ldots,j_r\in\{1:s\}} b_{j_1} a_{j_1 j_2} \times \ldots \times a_{j_{r-1} j_r} &= \sum_{j_2,\ldots,j_r\in\{1:s\}} b_{j_2}(1-c_{j_2}) a_{j_2 j_3} \times \ldots \times a_{j_{r-1} j_r} \\ &= \sum_{j_3,\ldots,j_r\in\{1:s\}} \frac{1}{2} b_{j_3}(1-c_{j_3})^2 a_{j_3 j_4} \times \ldots \times a_{j_{r-1} j_r} \\ &= \ldots = \sum_{j_r\in\{1:s\}} \frac{1}{(r-1)!} b_{j_r}(1-c_{j_r})^{r-1} = \frac{1}{r!},\end{aligned}$$

where the last identity follows from (78.9a) (actually the expression from Step (i) with $m := r - 1$ and $n := 0$; this is legitimate since $m + n + 1 \le r \le p$).
(v) Let us now show that (78.10) with $q := 1$ is met for all $r \in \{1{:}p\}$ if (78.9a) holds and (78.9b) and (78.9c) hold with $\eta + \zeta + 1 = p$. The case $r := 1$ has already been proved, and we have already established the result if either $r - 1 \le \eta$ or $r - 1 \le \zeta$. Let now $r \in \{2{:}p\}$ and assume that $r - 1 > \eta$ and $r - 1 > \zeta$ (i.e., $r - \eta \ge 2$ and $r - \zeta \ge 2$). Using (78.9b) for all $q \in \{1{:}\eta\}$, we obtain

$$\begin{aligned}\sum_{j_2,\ldots,j_r\in\{1:s\}} a_{j_1 j_2} \times \ldots \times a_{j_{r-1} j_r} &= \sum_{j_2,\ldots,j_{r-1}\in\{1:s\}} a_{j_1 j_2} \times \ldots \times a_{j_{r-2} j_{r-1}} \sum_{j_r\in\{1:s\}} a_{j_{r-1} j_r} \\ &= \ldots = \sum_{j_2,\ldots,j_{r-\eta}\in\{1:s\}} a_{j_1 j_2} \times \ldots \times a_{j_{r-\eta-1} j_{r-\eta}} \frac{1}{\eta!} c^\eta_{j_{r-\eta}}.\end{aligned}$$

Now, invoking (78.9c) (actually the expression from Step (ii) for all $m \in \{1{:}r{-}\eta{-}1\}$; this is legitimate since $r - \eta - 1 \le \zeta$ because $r \le p = \eta + \zeta + 1$) gives

$$\begin{aligned}&\sum_{j_1,\ldots,j_r\in\{1:s\}} b_{j_1} a_{j_1 j_2} \times \ldots \times a_{j_{r-\eta-1} j_{r-\eta}} \frac{1}{\eta!} c^\eta_{j_{r-\eta}} = \\ &\qquad \frac{1}{\eta!} \frac{1}{(r-\eta-1)!} \sum_{j_{r-\eta}\in\{1:s\}} b_{j_{r-\eta}}(1-c_{j_{r-\eta}})^{r-\eta-1} c^\eta_{j_{r-\eta}}.\end{aligned}$$

We now invoke the result from Step (i) with $m := r - \eta - 1$ and $n := \eta$. We infer that

$$\sum_{j_1,\ldots,j_r\in\{1:s\}} b_{j_1} a_{j_1 j_2} \times \ldots \times a_{j_{r-\eta-1} j_{r-\eta}} \frac{1}{\eta!} c^\eta_{j_{r-\eta}} = \frac{1}{\eta!} \frac{1}{(r-\eta-1)!} \frac{(r-\eta-1)!\eta!}{r!}.$$

In conclusion, we have proved that

$$\sum_{j_1,\ldots,j_r\in\{1:s\}} b_{j_1} a_{j_1 j_2} \times \ldots \times a_{j_{r-1} j_r} = \frac{1}{r!}, \qquad \forall r \in \{1{:}p\}.$$

Hence, (78.10) with $q := 1$ is met for all $r \in \{1{:}p\}$.
(vi) Let us finally show that (78.10) is met for all $r \in \{1{:}p\}$ and all $q \in \{1{:}p{-}r{+}1\}$ if (78.9a) holds and (78.9b) and (78.9c) hold with $p \le \eta + \zeta + 1$. Let $r \in \{1{:}p\}$. We are going to consider three cases: $r - 1 \le \zeta$, $1 \le \zeta \le r - 2$, and $\zeta = 0$.
Case 1: Assume that $r - 1 \le \zeta$. Using first Step (ii) for all $m \in \{1{:}r{-}1\}$, then Step (i) because

$r+q-2+1 \le r+p-r+1-1 = p$, we have

$$\begin{aligned}
&\sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1}a_{j_1j_2}\times\dots\times a_{j_{r-1}j_r}c_{j_r}^{q-1}\\
&\qquad=\sum_{j_2,\dots,j_{r-1}\in\{1:s\}} b_{j_2}(1-c_{j_2})a_{j_2j_3}\times\dots\times a_{j_{r-2}j_{r-1}}a_{j_{r-1}j_r}c_{j_r}^{q-1}\\
&\qquad=\dots=\frac{1}{(r-1)!}\sum_{j_r\in\{1:s\}} b_{j_r}(1-c_{j_r})^{r-1}c_{j_r}^{q-1}\\
&\qquad=\frac{1}{(r-1)!}\frac{(r-1)!(q-1)!}{(r+q-1)!}=\frac{(q-1)!}{(r+q-1)!}.
\end{aligned}$$

Case 2: Assume $1\le\zeta\le r-2$. Proceeding as above, we have

$$\begin{aligned}
&\sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1}a_{j_1j_2}\times\dots\times a_{j_{r-1}j_r}c_{j_r}^{q-1}\\
&\qquad=\sum_{j_2,\dots,j_{r-1}\in\{1:s\}} b_{j_2}(1-c_{j_2})a_{j_2j_3}\times\dots\times a_{j_{r-2}j_{r-1}}a_{j_{r-1}j_r}c_{j_r}^{q-1}\\
&\qquad=\dots=\frac{1}{(\zeta-1)!}\sum_{j_{\zeta+1},\dots,j_r\in\{1:s\}} b_{j_{\zeta+1}}(1-c_{j_{\zeta+1}})^{\zeta}a_{j_{\zeta+1}j_{\zeta+2}}\times\dots\times a_{j_{r-1}j_r}c_{j_r}^{q-1}.
\end{aligned}$$

Then, using (78.9b) (because $r-\zeta-1+q-1\le r-\zeta-2+p-r+1=p-\zeta-1\le\eta$), we obtain

$$\begin{aligned}
&\sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1}a_{j_1j_2}\times\dots\times a_{j_{r-1}j_r}c_{j_r}^{q-1}\\
&\qquad=\frac{1}{\zeta!}\sum_{j_{\zeta+1},\dots,j_r\in\{1:s\}} b_{j_{\zeta+1}}(1-c_{j_{\zeta+1}})^{\zeta}a_{j_{\zeta+1}j_{\zeta+2}}\times\dots\times\sum_{j_r\in\{1:s\}} a_{j_{r-1}j_r}c_{j_r}^{q-1}\\
&\qquad=\dots=\frac{1}{\zeta!}\sum_{j_{\zeta+1}\in\{1:s\}} b_{j_{\zeta+1}}(1-c_{j_{\zeta+1}})^{\zeta}\frac{(q-1)!}{(r-\zeta-1+q-1)!}c_{j_{\zeta+1}}^{r-\zeta-1+q-1}.
\end{aligned}$$

We now conclude by using Step (i) (because $\zeta+r-\zeta-1+q-1+1\le r-1+p-r+1=p$), which gives

$$\begin{aligned}
&\sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1}a_{j_1j_2}\times\dots\times a_{j_{r-1}j_r}c_{j_r}^{q-1}\\
&\qquad=\frac{1}{\zeta!}\frac{(q-1)!}{(r-\zeta-1+q-1)!}\frac{\zeta!(r-\zeta-1+q-1)!}{(r+q-1)!}=\frac{(q-1)!}{(r+q-1)!}.
\end{aligned}$$

Case 3: Assume $\zeta=0$. Proceeding as above and using (78.9b) (because $r-1+q-1\le r-2+$

$p-r+1=p-1\le\eta+\zeta=\eta$) then using (78.9a) (because $r-1+q-1\le p-1\le p$), we have

$$\begin{aligned}\sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1}a_{j_1j_2}\times\dots\times a_{j_{r-1}j_r}c_{j_r}^{q-1}&=\sum_{j_1,\dots,j_{r-1}\in\{1:s\}} b_{j_1}a_{j_1j_2}\times\dots\times\sum_{j_r\in\{1:s\}}a_{j_{r-1}j_r}c_{j_r}^{q-1}\\&=\dots=\sum_{j_1\in\{1:s\}}b_{j_1}\frac{(q-1)!}{(r-1+q-1)!}c_{j_1}^{r-1+q-1}=\frac{(q-1)!}{(r+q-2)!}\frac{1}{(r-1+q)}\\&=\frac{(q-1)!}{(r+q-1)!}.\end{aligned}$$

Exercise 78.3 (Explicit Euler). This time we use the identity $(u_h^n-u_h^{n-1},u_h^n)_L=\frac12\|u_h^n\|_L^2-\frac12\|u_h^{n-1}\|_L^2+\frac12\|u_h^n-u_h^{n-1}\|_L^2$ and we observe that $a_h(u_h^{n-1},u_h^n)=a_h(u_h^n,u_h^n)+a_h(u_h^{n-1}-u_h^n,u_h^n)$. Rearranging the terms, we infer that

$$\|u_h^n\|_L^2-\|u_h^{n-1}\|_L^2+\|u_h^n-u_h^{n-1}\|_L^2\le\frac{\tau}{\rho}\|u_h^n\|_L^2+\rho\tau\|\alpha^{n,1}\|_L^2+2\tau|a_h(u_h^n-u_h^{n-1},u_h^n)|.$$

Notice that the first term on the right-hand side is now $\frac{\tau}{\rho}\|u_h^n\|_L^2$ and no longer $\frac{\tau}{\rho}\|u_h^{n-1}\|_L^2$ owing to our choice of the test function. We need to bound $2\tau|a_h(u_h^n-u_h^{n-1},u_h^n)|$ on the right-hand side, and to this purpose, we can exploit the nonnegative term $\|u_h^n-u_h^{n-1}\|_L^2$ on the left-hand side. Invoking the estimate $|a_h(v_h,w_h)|\le\varpi\beta h^{-1}\|v_h\|_L\|w_h\|_L$ for all $v_h,w_h\in V_h$, we obtain

$$\begin{aligned}2\tau|a_h(u_h^n-u_h^{n-1},u_h^n)|&\le\tau\tfrac{\rho\beta^2}{\lambda_0h^2}\|u_h^n-u_h^{n-1}\|_L^2+\lambda_0\varpi^2\tfrac{\tau}{\rho}\|u_h^n\|_L^2\\&\le\|u_h^n-u_h^{n-1}\|_L^2+\lambda_0\varpi^2\tfrac{\tau}{\rho}\|u_h^n\|_L^2.\end{aligned}$$

Putting everything together yields

$$\|u_h^n\|_L^2-\|u_h^{n-1}\|_L^2\le(1+\lambda_0\varpi^2)\frac{\tau}{\rho}\|u_h^n\|_L^2+\rho\tau\|\alpha^{n,1}\|_L^2.$$

We conclude using the discrete Gronwall lemma from Exercise 68.3 with $\gamma:=(1+\lambda_0\varpi^2)\frac{\tau}{\rho}\le\frac12$ by assumption and using that $\frac{1}{1-\gamma}\le e^{2\gamma}$.

Exercise 78.4 (First-order viscosity). (i) The explicit scheme only requires solving Hermitian positive definite linear systems, whereas the implicit Euler scheme involves a non-Hermitian linear system (think of A being the transport operator $\boldsymbol{\beta}\cdot\nabla u$).
(ii) Let us test the equation with τu_h^{n-1}. We obtain

$$\frac12\|u_h^n\|_L^2-\frac12\|u_h^{n-1}\|_L^2-\frac12\|u_h^n-u_h^{n-1}\|_L^2+\tau(A(u_h^{n-1}),u_h^{n-1})_L+\tau\mu|u_h^{n-1}|_V^2=0.$$

Taking the real part and multiplying by 2 gives

$$\|u_h^n\|_L^2-\|u_h^{n-1}\|_L^2+2\tau\mu|u_h^{n-1}|_V^2\le\|u_h^n-u_h^{n-1}\|_L^2.$$

Moreover, testing the equation with $\tau(u_h^n-u_h^{n-1})$ gives

$$\begin{aligned}\|u_h^n-u_h^{n-1}\|_L^2&\le\tau\|A(u_h^{n-1})\|_L\|u_h^n-u_h^{n-1}\|_L+\tau\mu|u_h^{n-1}|_V|u_h^n-u_h^{n-1}|_V\\&\le\tau\beta|u_h^{n-1}|_V\|u_h^n-u_h^{n-1}\|_L+\tau\mu|u_h^{n-1}|_V|u_h^n-u_h^{n-1}|_V\\&\le\tau\beta|u_h^{n-1}|_V\|u_h^n-u_h^{n-1}\|_L+\tau\mu c_{\mathrm{INV}}(h)|u_h^{n-1}|_V\|u_h^n-u_h^{n-1}\|_L.\end{aligned}$$

Hence, we have

$$\|u_h^n - u_h^{n-1}\|_L \le \tau(\beta + \mu c_{\mathrm{INV}}(h))|u_h^{n-1}|_V.$$

We infer that

$$\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 + 2\tau\mu|u_h^{n-1}|_V^2 \le \tau^2(\beta + \mu c_{\mathrm{INV}}(h))^2|u_h^{n-1}|_V^2.$$

Then, provided $\tau(\beta + \mu c_{\mathrm{INV}}(h))^2 \le 2\mu$, we have $\|u_h^n\|_L \le \|u_h^{n-1}\|_L$, which readily implies that $\|u_h^n\|_L \le \|u_h^0\|_L$ for all $n \in \mathcal{N}_\tau$.
(iii) The stability condition is equivalent to

$$\tau(\beta^2 + 2\mu\beta c_{\mathrm{INV}}(h) + \mu^2 c_{\mathrm{INV}}(h)^2) - 2\mu = \mu^2(c_{\mathrm{INV}}(h)^2\tau) + 2\mu(\beta c_{\mathrm{INV}}(h)\tau - 1) + \beta^2\tau \le 0.$$

The above quadratic function in μ can take negative values if and only if the discriminant is nonnegative, which gives $\beta\tau c_{\mathrm{INV}}(h) \le \frac{1}{2}$. Let $\lambda := \beta\tau c_{\mathrm{INV}}(h)$ be the CFL number. Thus, the admissible range for μ is

$$\frac{\beta}{c_{\mathrm{INV}}(h)}\frac{1-\lambda-\sqrt{1-2\lambda}}{\lambda} \le \mu \le \frac{\beta}{c_{\mathrm{INV}}(h)}\frac{1-\lambda+\sqrt{1-2\lambda}}{\lambda}.$$

Exercise 78.5 (Explicit Euler, mass lumping). (i) A direct computation shows that for $n \in \mathcal{N}_\tau$,

$$h(\mathsf{U}_j^n - \mathsf{U}_j^{n-1}) + \frac{\tau\beta}{2}(\mathsf{U}_{j+1}^{n-1} - \mathsf{U}_{j-1}^{n-1}) = 0,$$

for all $j \in \{0{:}I\}$.
(ii) The above computation gives

$$\mathsf{U}_j^n = \mathsf{U}_j^{n-1} - \frac{\beta\tau}{2h}(\mathsf{U}_{j+1}^{n-1} - \mathsf{U}_{j-1}^{n-1}).$$

Let us set $\lambda := \frac{\beta\tau}{2h}$ and square the above equation. This yields

$$(\mathsf{U}_j^n)^2 = (\mathsf{U}_j^{n-1})^2 + \lambda^2(\mathsf{U}_{j+1}^{n-1} - \mathsf{U}_{j-1}^{n-1})^2 - 2\lambda\mathsf{U}_j^{n-1}\mathsf{U}_{j+1}^{n-1} + 2\lambda\mathsf{U}_j^{n-1}\mathsf{U}_{j-1}^{n-1}.$$

Summing over j and using that $\sum_{j\in\{1:I\}}\mathsf{U}_j^{n-1}\mathsf{U}_{j+1}^{n-1} = \sum_{j\in\{1:I\}}\lambda\mathsf{U}_j^{n-1}\mathsf{U}_{j-1}^{n-1}$ owing to the periodic boundary conditions, we obtain

$$\sum_{j\in\{1:I\}}(\mathsf{U}_j^n)^2 = \sum_{j\in\{1:I\}}(\mathsf{U}_j^{n-1})^2 + \lambda^2\sum_{j\in\{1:I\}}(\mathsf{U}_{j+1}^{n-1} - \mathsf{U}_{j-1}^{n-1})^2.$$

(iii) Let us prove by induction that $\mathsf{U}_j^n = a^{n+1}e^{\mathrm{i}\frac{j}{I}2k\pi}$ for all $j \in \{1{:}I\}$. This is true for $n = 0$. By definition, we have

$$\begin{aligned}
\mathsf{U}_j^n &= \mathsf{U}_j^{n-1} - \lambda(\mathsf{U}_{j+1}^{n-1} - \mathsf{U}_{j-1}^{n-1}) \\
&= a^n e^{\mathrm{i}\frac{j}{I}2k\pi} - \lambda a^n(e^{\mathrm{i}\frac{j+1}{I}2k\pi} - e^{\mathrm{i}\frac{j-1}{I}2k\pi}) \\
&= a^n e^{\mathrm{i}\frac{j}{I}2k\pi}\big(1 - \lambda(e^{\mathrm{i}\frac{k}{I}2\pi} - e^{-\mathrm{i}\frac{k}{I}2k\pi})\big) \\
&= a^n e^{\mathrm{i}\frac{j}{I}2k\pi}\big((1 - 2\mathrm{i}\lambda\sin(\tfrac{k}{I}2\pi)\big) = a^{n+1}e^{\mathrm{i}\frac{j}{I}2k\pi}.
\end{aligned}$$

This proves the assertion. Finally, since $|a| = \big(1 + 4\lambda^2(\sin(\frac{k}{I}2\pi))^2\big)^{\frac{1}{2}} > 1$, we conclude that $|\mathsf{U}_j^n| = |a|^{n+1} \to \infty$ as $n \to \infty$ for all $j \in \{1{:}I\}$. Thus, the explicit Euler scheme with mass lumping is unstable.

Exercise 78.6 (Error equation, RK2). (i) Integrating by parts in time, we obtain

$$\begin{aligned}\tfrac{1}{2}\tau\psi^{n-1} &= \int_{J_n} \tfrac{1}{2}(t-t_n)^2\partial_{ttt}u(t)\mathrm{d}t\\ &= -\int_{J_n}(t-t_n)\partial_{tt}u(t)\mathrm{d}t - \tfrac{1}{2}\tau^2\partial_{tt}u(t_{n-1})\\ &= \int_{J_n}\partial_t u(t)\mathrm{d}t - \tau\partial_t u(t_{n-1}) - \tfrac{1}{2}\tau^2\partial_{tt}u(t_{n-1})\\ &= u(t_n) - u(t_{n-1}) - \tau\partial_t u(t_{n-1}) - \tfrac{1}{2}\tau^2\partial_{tt}u(t_{n-1}).\end{aligned}$$

This proves the assertion.
(ii) Using the result from Step (i), we obtain

$$\begin{aligned}&(u(t_n) - \tfrac{1}{2}(y(t_{n-1}) + u(t_{n-1})), w_h)_L\\ &= \tfrac{1}{2}\tau(\partial_t u(t_{n-1}), w_h)_L + \tfrac{1}{2}\tau^2(\partial_{tt}u(t_{n-1}), w_h)_L + \tfrac{1}{2}\tau(\psi^{n-1}, w_h)_L,\end{aligned}$$

for all $w_h \in V_h$. Moreover, we have

$$\begin{aligned}(\partial_t u(t_{n-1}), w_h)_L + (A(u(t_{n-1})), w_h)_L &= (f^{n-1}, w_h)_L,\\ (\partial_{tt}u(t_{n-1}), w_h)_L + (A(\partial_t u(t_{n-1})), w_h)_L &= (\partial_t f^{n-1}, w_h)_L.\end{aligned}$$

Hence, we have

$$(\partial_t u(t_{n-1}), w_h)_L + \tau(\partial_{tt}u(t_{n-1}), w_h)_L + (A(y(t_{n-1})), w_h)_L = (f^{n-1} + \tau\partial_t f^{n-1}, w_h)_L.$$

Putting everything together yields

$$\begin{aligned}&(u(t_n) - \tfrac{1}{2}(y(t_{n-1}) + u(t_{n-1})), w_h)_L + \tfrac{1}{2}\tau(A(y(t_{n-1})), w_h)_L\\ &\quad= \tfrac{1}{2}\tau(\partial_t u(t_{n-1}), w_h)_L + \tfrac{1}{2}\tau^2(\partial_{tt}u(t_{n-1}), w_h)_L + (A(y(t_{n-1})), w_h)_L + \tfrac{1}{2}\tau(\psi^{n-1}, w_h)_L\\ &\quad= \tfrac{1}{2}\tau(f^{n-1} + \tau\partial_t f^{n-1}, w_h)_L + \tfrac{1}{2}\tau(\psi^{n-1}, w_h)_L.\end{aligned}$$

This completes the proof of (78.26).

Exercise 78.7 (ERK schemes, $p = 3$). (i) Let $\widetilde{u}_{h\tau}$ be the sequence produced by (78.27) with $\alpha^{n,3}$ replaced by $\widetilde{\alpha}^{n,3} := \alpha^{n,3} + r_h^{n,3}$. Eliminating the intermediate stages in (78.27), we obtain

$$\widetilde{u}_h^n = (I_{V_h} - \tau A_h + \tfrac{1}{2}\tau^2 A_h^2 - \tfrac{1}{6}\tau^3 A_h^3)(\widetilde{u}_h^{n-1}) + \tau\big(\mathsf{G}_3(t_{n-1}) + \tfrac{1}{3}r_h^{n,3}\big),$$

where

$$\mathsf{G}_3(t_{n-1}) := f_h^{n-1} + \tfrac{1}{2}\tau\big(\partial_t f_h^{n-1} - A_h(f_h^{n-1})\big) + \tfrac{1}{6}\tau^2\big(\partial_{tt}f_h^{n-1} - A_h(\partial_t f_h^{n-1}) + A_h^2(f_h^{n-1})\big).$$

(If $r_h^{n,3} = 0$, $\widetilde{u}_h^n$ exactly reproduces the third-order Taylor expansion of the solution $u_h(t)$ at t_n; see (78.1) and (78.13) with $p := 3$). Moreover, eliminating the intermediate stages in the ERK scheme leads to

$$u_h^n = (I_{V_h} - \tau A_h + \tfrac{1}{2}\tau^2 A_h^2 - \tfrac{1}{6}\tau^3 A_h^3)(u_h^{n-1}) + \tau\Delta_3^n,$$

where

$$\begin{aligned}\Delta_3^n &:= b_1 f_h(t_{n,1}) + b_2 f_h(t_{n,2}) + b_3 f_h(t_{n,3})\\ &\quad - \tau A_h\big((b_2 a_{21} + b_3 a_{31}) f_h(t_{n,1}) + b_3 a_{32} f_h(t_{n,2})\big) + \tau^2\tfrac{1}{6}A_h^2\big(f_h(t_{n,1})\big),\end{aligned}$$

and we used $b_1 + b_2 + b_3 = 1$, $b_2a_{21} + b_3a_{31} + b_3a_{32} = \frac{1}{2}$, $b_3a_{32}a_{21} = \frac{1}{6}$ (see Example 78.11). An induction argument shows that the sequences $u_{h\tau}$ and $\widetilde{u}_{h\tau}$ coincide if $r_h^{n,3}$ is chosen so that $r_h^{n,3} := 3(\Delta_3^n - \mathsf{G}_3(t_{n-1}))$. This is equivalent to setting $r_h^{n,3} := 3\big(\mathcal{P}_{V_h}(r_1^{n,3}) - \tau A_h(\mathcal{P}_{V_h}(r_2^{n,3}))\big)$ with

$$\begin{aligned}
r_1^{n,3} &:= b_1 f(t_{n,1}) + b_2 f(t_{n,2}) + b_3 f(t_{n,3}) - f^{n-1} - \tfrac{1}{2}\tau\partial_t f^{n-1} - \tfrac{1}{6}\tau^2\partial_{tt} f^{n-1},\\
r_2^{n,3} &:= (b_2a_{21} + b_3a_{31}) f(t_{n,1}) + b_3a_{32} f(t_{n,2}) - \tfrac{1}{2} f^{n-1} - \tfrac{1}{6}\tau\partial_t f^{n-1}.
\end{aligned}$$

This proves the assertion (i) in Lemma 78.21.
(ii) Let us now prove the assertion (ii). Using that $\sum_{j\in\{1:3\}} b_j c_j^{q-1} = \frac{1}{q}$ for all $q \in \{1{:}3\}$, we infer that

$$\begin{aligned}
r_1^{n,3} &= \sum_{j\in\{1:3\}} b_j \left(f(t_{n,j}) - f^{n-1} - c_j\tau\partial_t f^{n-1} - \tfrac{1}{2}c_j^2\tau^2\partial_{tt} f^{n-1}\right)\\
&= \sum_{j\in\{1:3\}} \frac{1}{2} b_j \int_{t_{n-1}}^{t_{n,j}} (t_{n,j} - t)^2 \partial_{ttt} f(t)\mathrm{d}t.
\end{aligned}$$

Moreover, using that $b_2a_{21} + b_3a_{31} + b_3a_{32} = \frac{1}{2}$, $b_2a_{21}c_1 + b_3a_{31}c_1 + b_3a_{32}c_2 = \frac{1}{6}$, we obtain

$$\begin{aligned}
r_2^{n,3} &= (b_2a_{21} + b_3a_{31}) \int_{t_{n-1}}^{t_{n,1}} \partial_t f\mathrm{d}t + b_3a_{32} \int_{t_{n-1}}^{t_{n,2}} \partial_t f\mathrm{d}t - \tfrac{1}{6}\tau\partial_t f^{n-1}\\
&= (b_2a_{21} + b_3a_{31}) \int_{t_{n-1}}^{t_{n,1}} (t_{n,1} - t)\partial_{tt} f\mathrm{d}t + b_3a_{32} \int_{t_{n-1}}^{t_{n,2}} (t_{n,2} - t)\partial_{tt} f\mathrm{d}t.
\end{aligned}$$

(Notice that altogether we used all the necessary order conditions from Example 78.11.) Using the L-stability of $\mathcal{P}_{V_h}$, the bound $\|A_h(w_h)\|_L \le c\frac{1}{\rho}\|w_h\|_{H^1}$ for all $w_h \in V_h$ (recall that $\rho\beta \le \ell_D$ by definition of ρ), and the H^1-stability of $\mathcal{P}_{V_h}$ (recall that the mesh sequence is quasi-uniform and invoke Proposition 22.21), we infer that

$$\begin{aligned}
\|r_h^{n,3}\|_L &\le 3\|\mathcal{P}_{V_h}(r_1^{n,3})\|_L + 3\tau\|A_h(\mathcal{P}_{V_h}(r_2^{n,3}))\|_L\\
&\le 3\|r_1^{n,3}\|_L + c\,\tau\rho^{-1}\|r_2^{n,3}\|_{H^1(D;\mathbb{C}^m)}.
\end{aligned}$$

We can then conclude from the above identities that $r_h^{n,3}$ satisfies (78.28).

Chapter 79

Scalar conservation equations

Exercises

Exercise 79.1 (Kružkov entropy pairs). For all $k \in \mathbb{R}$, consider the entropy $\eta(v,k) := |v-k|$. Compute the entropy flux associated with this entropy, $\boldsymbol{q}(v)$, with the normalization $\boldsymbol{q}(k) := \mathbf{0}$.

Exercise 79.2 (Entropy solution). Consider Burgers' equation with $D := \mathbb{R}$ and $u_0(x) := 0$. (i) What should be the entropy solution to this problem? (ii) Let H be the Heaviside function. Let $a \in \mathbb{R}$ and consider $u(x,t) := 2aH(x) - aH(x - \frac{at}{2}) - aH(x + \frac{at}{2})$. Draw the graph of $u(\cdot,t)$ at some time $t > 0$. (iii) Show that u is a weak solution for all $a \in \mathbb{R}$. (iv) Verify that u is not the entropy solution. (*Hint*: consider the entropy $\eta(v) := |v|$.)

Exercise 79.3 (Entropy solution). Consider Burgers' equation with $D := \mathbb{R}$ and $u_0(x) := H(x)$, where H is the Heaviside function. (i) Verify that $u_1(x,t) := H(x - \frac{1}{2}t)$ and $u_2(x,t) := 0$ if $x < 0$, $u_2(x,t) := \frac{x}{t}$, if $0 < x < t$, $u_2(x,t) := 1$ if $x > t$, are both weak solutions. (ii) Verify that u_1 does not satisfy the entropy inequalities, whereas u_2 does.

Exercise 79.4 (Average speed). Let f be a scalar Lipschitz flux. Consider the Riemann problem $\partial_t u + \partial_x f(u) = 0$, with initial data (u_L, u_R), $u_L \neq u_R$. Let $\lambda_{\max}(u_L, u_R)$ be a maximum wave speed in this problem. Let $s := (f(u_L) - f(u_R))/(u_L - u_R)$ be the average speed. Assume that the interval $[u_L, u_R]$ can be divided into finitely many intervals where f has a continuous and bounded second derivative and f is either strictly convex or strictly concave. Prove that $|\lambda_{\max}(u_L, u_R)| \geq |s|$.

Exercise 79.5 (Maximum speed). Compute $\lambda_{\max}(u_L, u_R)$ for the two cases $(u_L, u_R) := (1, 2)$ and $(u_L, u_R) := (2, 1)$ with the following fluxes: (i) $f(v) := \frac{1}{2}v^2$; (ii) $f(v) := 8(v - \frac{1}{2})^3$; (iii) $f(v) := -(v-1)(2v-3)$ if $v \leq \frac{3}{2}$ and $f(v) := \frac{1}{4}(3 - 2v)$ if $\frac{3}{2} \leq v$.

Exercise 79.6 (Strong solutions). The goal is to justify Remark 79.13. (i) Show that if u is a weak solution and $u \in C^1(D \times [0, T^*))$, then u is a strong solution in $D \times [0, T^*)$. (ii) Show that if u is a strong solution, then u is also a weak solution. (iii) Let u be a strong solution to (79.1) and let $(\eta, \boldsymbol{q})$ an entropy pair with η of class C^2. Show that (79.10) holds true.

Exercise 79.7 (Method of characteristics). Let $D := \mathbb{R}$, $\boldsymbol{f} := f\boldsymbol{e}_x$, and assume that f is of class C^2 and u_0 is of class C^1. Recall that there exists $T^* > 0$ and a unique $s(x,t)$ solving $x = f'(u_0(s))t + s$ for all x and all $t \in [0, T^*)$. (i) Show that $u(x,t) := u_0(s(x,t))$ solves (79.1)

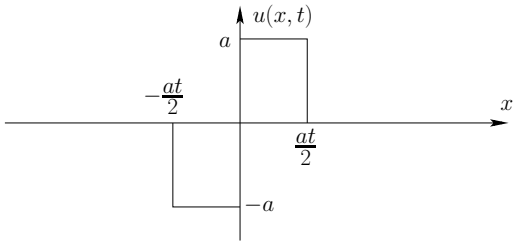
for all $t \in [0, T^*)$. (ii) Let $s_0 \in \mathbb{R}$. Show that $u(x,t)$ is constant along the straight segment $\{x = f'(u_0(s_0))t + s_0 \mid t \in [0, T^*]\}$. (iii) Show that the solution found in Step (i) is the entropy solution.

Exercise 79.8 (Shock interacting with an expansion wave). Consider Burgers' equation with the initial condition $u_0(x) := -1$ if $x \in (-1, 0)$ and $u_0(x) := 0$ otherwise. (i) Derive the weak entropy solution up to the time $t = 2$. (ii) After the time $t = 2$, the shock originating from $x = -1$ starts interacting with the expansion wave originating from $x = 0$, leading to a shock with a nonlinear trajectory. Derive the weak entropy solution for the times $t \geq 2$. (*Hint*: use the Rankine–Hugoniot condition.) (iii) Verify that "mass" conservation is satisfied, i.e., $\int_{\mathbb{R}} u(x,t)\,\mathrm{d}x = \int_{\mathbb{R}} u_0(x)\,\mathrm{d}x = -1$ for all $t \geq 0$.

Solution to exercises

Exercise 79.1 (Kružkov entropy pairs). By definition, $\boldsymbol{q}(u) = \int_k^u \operatorname{sign}(v-k)\boldsymbol{f}'(v)\,\mathrm{d}v$, where $\operatorname{sign}(z) = -1$ if $z < 0$ and $\operatorname{sign}(z) = 1$ if $z > 0$. If $u < k$, then $\boldsymbol{q}(u) = -\int_k^u \boldsymbol{f}'(v)\,\mathrm{d}v = \int_u^k \boldsymbol{f}'(v)\,\mathrm{d}v = \boldsymbol{f}(k) - \boldsymbol{f}(u) = \operatorname{sign}(u-k)(\boldsymbol{f}(u) - \boldsymbol{f}(k))$. We obtain the same result if $k < u$. Hence, $\boldsymbol{q}(u) = \operatorname{sign}(u-k)(\boldsymbol{f}(u) - \boldsymbol{f}(k))$.

Exercise 79.2 (Entropy solution). Recall that Burgers' flux is $f(v) = \frac{1}{2}v^2$ and Burgers' equation is $\partial_t v + \frac{1}{2}\partial_x v^2 = 0$.
(i) $u(x,t) = 0$ is clearly a weak solution to this problem. It also trivially satisfies all the entropy inequalities. Hence, it is the entropy solution.
(ii) Here is the graph of $u(\cdot, t)$ at some $t > 0$ with $a > 0$:



(iii) One possibility is to verify that $\partial_t u + \partial_x(\frac{1}{2}u^2) = 0$ is satisfied in the sense of distributions. Let us consider $u(x,t) = 2aH(x) - aH(x - \frac{at}{2}) - aH(x + \frac{at}{2})$. For all $\alpha \in \mathbb{R}$, let $\Sigma(\alpha) := \{(x,t) \in \mathbb{R}\times\mathbb{R}_+ \mid x = \alpha t\}$, and let $\delta_{\Sigma(\alpha)}$ be the line Dirac measure defined s.t. $\langle \delta_{\Sigma(\alpha)}, \phi\rangle := \int_{\mathbb{R}_+} \phi(\alpha s, s)\,\mathrm{d}s$ for all $\phi \in C_0^\infty(\mathbb{R}\times\mathbb{R}_+)$. We have $\partial_t H(x - \alpha t) = -\alpha\delta_{\Sigma(\alpha)}$, so that

$$\begin{aligned}\partial_t u &= \partial_t\Big(2aH(x) - aH(x - \frac{at}{2}) - aH(x + \frac{at}{2})\Big)\\ &= \frac{1}{2}a^2\delta_{\Sigma(\frac{a}{2})} - \frac{1}{2}a^2\delta_{\Sigma(-\frac{a}{2})}.\end{aligned}$$

Moreover, we have

$$\begin{aligned} u^2(x,t) &= 4a^2H(x) + a^2H(x-\frac{at}{2}) + a^2H(x+\frac{at}{2}) \\ &\quad - 4a^2H(x-\frac{at}{2}) - 4a^2H(x) + 2a^2H(x-\frac{at}{2}) \\ &= -a^2H(x-\frac{at}{2}) + a^2H(x+\frac{at}{2}), \end{aligned}$$

where we used that $H(y_1)H(y_2) = H(\max(y_1,y_2))$. We infer that $\partial_x(\frac{1}{2}u^2) = -\frac{1}{2}a^2\delta_{\Sigma(\frac{a}{2})} + \frac{1}{2}a^2\delta_{\Sigma(-\frac{a}{2})}$. Hence, $\partial_t u + \partial_x(\frac{1}{2}u^2) = 0$, thereby proving that u is a weak solution.

Another possibility is to verify that (79.7) holds true for all $\phi \in C_0^1(\mathbb{R}\times\mathbb{R}_+)$. We have

$$\begin{aligned} \int_0^\infty \int_{\mathbb{R}} u\partial_t\phi \,\mathrm{d}x\mathrm{d}t &= \int_{\mathbb{R}}\int_0^\infty u\partial_t\phi \mathrm{d}t\,\mathrm{d}x \\ &= -a\int_{\mathbb{R}_-}\int_{-\frac{2x}{a}}^\infty \partial_t\phi \mathrm{d}t\,\mathrm{d}x + a\int_{\mathbb{R}_+}\int_{\frac{2x}{a}}^\infty \partial_t\phi \mathrm{d}t\,\mathrm{d}x \\ &= a\int_{\mathbb{R}_-}\phi(x,-\tfrac{2x}{a})\,\mathrm{d}x - a\int_{\mathbb{R}_+}\phi(x,\tfrac{2x}{a})\,\mathrm{d}x \\ &= \frac{1}{2}a^2\int_{\mathbb{R}_+}\phi(-\tfrac{as}{2},s)\,\mathrm{d}s - \frac{1}{2}a^2\int_{\mathbb{R}_+}\phi(\tfrac{as}{2},s)\,\mathrm{d}s. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \int_0^\infty \int_{\mathbb{R}} u^2\partial_x\phi \,\mathrm{d}x\mathrm{d}t &= a^2\int_{\mathbb{R}_+}\int_{-\frac{at}{2}}^{\frac{at}{2}} \partial_x\phi \,\mathrm{d}x\mathrm{d}t \\ &= a^2\int_{\mathbb{R}_+}\big(\phi(\tfrac{at}{2},t) - \phi(-\tfrac{at}{2},t)\big)\mathrm{d}t, \end{aligned}$$

which shows that $\int_0^\infty\int_{\mathbb{R}}\big(u\partial_t\phi + \frac{1}{2}u^2\partial_x\phi\big)\,\mathrm{d}x\mathrm{d}t = 0$.

(iv) One way to answer this question is to invoke Theorem 79.10. Since $u(x,t) = 0$ is an entropy solution and the entropy solution is unique, $u(x,t) = 2aH(x) - aH(x-\frac{at}{2}) - aH(x+\frac{at}{2})$ cannot be an entropy solution. Another way to answer the question is to exhibit one entropy such that the corresponding entropy inequality is violated. Let us take $\eta(v) := |v|$. Then, the corresponding entropy flux is $q(v) = \int_0^v \mathrm{sgn}(\xi)\xi\,\mathrm{d}\xi = \frac{1}{2}|v|v$. We have $|u(x,t)| = -aH(x-\frac{at}{2}) + aH(x+\frac{at}{2})$. Since $H(y)H(z) = H(z)$ for $y \le z$, we obtain

$$|u(x,t)|u(x,t) = 2a^2H(x) - a^2H(x-\tfrac{at}{2}) - a^2H(x+\tfrac{at}{2}) = au(x,t).$$

We infer that

$$\partial_t\eta(u) + \partial_x q(u) = \frac{1}{2}a^2\delta_{\frac{a}{2}} + \frac{1}{2}a^2\delta_{-\frac{a}{2}} + a^2\delta_0 - \frac{1}{2}a^2\delta_{\frac{a}{2}} - \frac{1}{2}a^2\delta_{-\frac{a}{2}} = a^2\delta_0.$$

But $\partial_t\eta(u) + \partial_x q(u) = a^2\delta_0$ is a positive distribution. Hence, the entropy inequality is violated. (Notice that $\int_0^\infty\int_{\mathbb{R}}(-\eta(u)\partial_t\phi(x,t) - q(u)\partial_x\phi(x,t))\,\mathrm{d}x\mathrm{d}t = a^2\int_0^\infty \phi(0,t)\mathrm{d}t$ for every smooth function compactly supported in $\mathbb{R}\times\mathbb{R}_+$.)

Exercise 79.3 (Entropy solution). (i) Let us look at u_1 first. In the distribution sense, we have $\partial_t u_1 = -\frac{1}{2}\delta(x-\frac{1}{2}t)$, where δ is the Dirac measure, and upon observing that $u_1^2 = u_1$, we also

have $\partial_x(\frac{1}{2}u_1^2) = -\frac{1}{2}\delta(x-\frac{1}{2}t)$. Hence, $\partial_t u_1 + \partial_x(\frac{1}{2}u_1^2) = 0$. Let us now look at u_2. Upon observing that u_2 is a continuous function in $\mathbb{R}\times\mathbb{R}_+$, we have $\partial_t u_2(x,t) = 0$ if $x < 0$, $\partial_t u_2(x,t) = -\frac{x}{t^2}$, if $0 < x < t$, $\partial_t u_2(x,0) = 0$ if $x > t$, and $\partial_x(\frac{1}{2}u_2^2(x,t)) = 0$ if $x < 0$, $\partial_x(\frac{1}{2}u_2^2(x,t)) = \frac{x}{t^2}$, if $0 < x < t$, $\partial_t(\frac{1}{2}u_2^2(x,0)) = 0$ if $x > t$. This proves that $\partial_t u_2 + \partial_x(\frac{1}{2}u_2^2) = 0$.
(ii) Let $k \in (0,1)$. Let us consider the Kružkov entropy pair $\eta(u) = |u-k|$ and $q(u) = \operatorname{sign}(u-k)(f(u)-f(k)) = \operatorname{sign}(u-k)\frac{1}{2}(u^2-k^2)$ (i.e., $\boldsymbol{q}(u) = q(u)\boldsymbol{e}_x$). Then, for u_1, we have $\eta(u) = |H(x-\frac{1}{2}t)-k| = k$ if $x < \frac{1}{2}t$ and $\eta(u_1) = |H(x-\frac{1}{2}t)-k| = 1-k$ if $x > \frac{1}{2}t$. This means that $\eta(u_1) = H(x-\frac{1}{2}t) - (2H(x-\frac{1}{2}t)-1)k$. This shows that $\partial_t\eta(u_1) = (-\frac{1}{2}+k)\delta(x-\frac{1}{2}t)$. Similarly, we have

$$q(u_1) = \operatorname{sign}(u_1-k)\frac{1}{2}(u_1^2-k^2) = \tfrac{1}{2}\operatorname{sign}(H(x-\tfrac{1}{2}t)-k)(H^2(x-\tfrac{1}{2}t)-k^2)$$
$$= \begin{cases} \frac{1}{2}k^2 & \text{if } x < \frac{1}{2}t, \\ \frac{1}{2}(1-k^2) & \text{if } x > \frac{1}{2}t. \end{cases}$$

This means that $\partial_x q(u_1) = (\frac{1}{2}-k^2)\delta(x-\frac{1}{2}t)\boldsymbol{e}_x$. Hence, we have

$$\partial_t\eta(u_1) + \partial_x q(u_1) = \big(-\tfrac{1}{2}+k+\tfrac{1}{2}-k^2\big)\delta(x-\tfrac{1}{2}t) = k(1-k)\delta(x-\tfrac{1}{2}t),$$

which proves that u_1 is not the entropy solution since $k(1-k)\delta(x-\frac{1}{2}t)$ is a positive measure for all $k \in (0,1)$.

We now do the computation for u_2. Clearly, $\partial_t\eta(u_2) + \partial_x q(u_2) = 0$ if $x < 0$ or $t < x$. Let us now assume that $0 < x < t$. Then $\eta(u_2) = |\frac{x}{t}-k| = k - \frac{x}{t}$ if $x < kt$ and $\eta(u_2) = |\frac{x}{t}-k| = \frac{x}{t}-k$ if $x > kt$, meaning that $\partial_t\eta(u_2) = +\frac{x}{t^2}$ if $x < kt$ and $\eta(u_2) = -\frac{x}{t^2}$ if $x > kt$. Similarly, $q(u_2) = \operatorname{sign}(u_2-k)\frac{1}{2}(u_2^2-k^2) = -\frac{1}{2}(\frac{x^2}{t^2}-k^2)$ if $x < kt$, and $q(u_2) = \frac{1}{2}(\frac{x^2}{t^2}-k^2)$ if $x > kt$, meaning that $\partial_x q(u_2) = -\frac{x}{t^2}$ if $x < kt$, and $q(u_2) = \frac{x}{t^2}$ if $x > kt$. Hence, $\partial_t\eta(u_2) + \partial_x q(u_2) = 0$ a.e. in x and t. In conclusion, u_2 is the entropy solution.

Exercise 79.4 (Average speed). From Theorem 79.15, we know that

$$\lambda_{\max}(u_L,u_R) \geq \max(|\underset{\sim}{f}'(u_L)|, |\underset{\sim}{f}'(u_R)|), \qquad \text{if } u_L < u_R,$$
$$\lambda_{\max}(u_L,u_R) \geq \max(|\bar{f}'(u_L)|, |\bar{f}'(u_R)|), \qquad \text{if } u_L < u_R.$$

Assume that $u_L < u_R$. Recall that $\underset{\sim}{f}(u_L) = f(u_L)$ and $\underset{\sim}{f}(u_R) = f(u_R)$. Hence, we have

$$\int_{u_L}^{u_R} \underset{\sim}{f}'(v)\,\mathrm{d}v = f(u_R) - f(u_L).$$

Since $\underset{\sim}{f}$ is convex, $\underset{\sim}{f}'$ is an increasing function, and we infer that

$$\underset{\sim}{f}'(u_R)(u_R-u_L) \geq \int_{u_L}^{u_R} \underset{\sim}{f}'(v)\,\mathrm{d}v = f(u_R) - f(u_L) \geq \underset{\sim}{f}'(u_L)(u_R-u_L).$$

This proves that $|s| \leq |\lambda_{\max}(u_L,u_R)|$. Notice in passing that $\underset{\sim}{f}'$ is continuous in the neighborhood of u_L and u_R since, by assumption, either $\underset{\sim}{f}$ is locally affine or $\underset{\sim}{f}$ is locally equal to f and f is of class C^2 in the neighborhood of u_L and u_R. This argument shows that the quantities $\underset{\sim}{f}'(u_R)$ and $\underset{\sim}{f}'(u_L)$ are well defined and are bounded. Similarly, if $u_L > u_R$, we have

$$\bar{f}'(u_R)(u_R-u_L) \leq \int_{u_L}^{u_R} \bar{f}'(v)\,\mathrm{d}v = f(u_R) - f(u_L) \geq \bar{f}'(u_L)(u_R-u_L),$$

and again $|s| \le |\lambda_{\max}(u_L, u_R)|$.

Exercise 79.5 (Maximum speed). (i) For $f(v) = \frac{1}{2}v^2$ we have: (a) $u_L < u_R$ and $\underset{\sim}{f} = f$, so that $\lambda_{\max}(1,2) = \max(|f'(1)|, |f'(2)|) = 2$; (b) in this case, the graph of $\overline{f}$ is the line connecting $(2, f(2))$ to $(1, f(1))$, so that $\lambda_{\max}(2,1) = \frac{1}{2}\frac{4-1}{2-1} = \frac{3}{2}$.
(ii) Assume now that $f(v) = 8(v - \frac{1}{2})^3 = (2v-1)^3$. For case (a), we have $u_L < u_R$ and f is convex on the interval $[1,2]$, so that we have $\underset{\sim}{f} = f_{|[1,2]}$. This implies that $\lambda_{\max}(1,2) = \max(|f'(u_L)|, |f'(u_R)|) = f'(u_R) = 6$; (b) $u_L > u_R$ but $\overline{f} \neq f$. Over the interval $[1,2]$, the graph of $\overline{f}$ is a straight line, so that $\lambda_{\max}(1,2) = |(f(u_L) - f(u_R))/(u_L - u_R)| = 26$.
(iii) For case (a), we have $u_L < u_R$ but $\underset{\sim}{f} \neq f$. In this case , the graph of $\underset{\sim}{f}$ is the line connecting $(1, f(1))$ to $(2, f(2))$, so that $\lambda_{\max}(1,2) = |(f(u_L) - f(u_R))/(u_L - u_R)| = |(0 - (-\frac{1}{4}))/(1-2)| = \frac{1}{4}$. In case (b), $\overline{f} \neq f$. Notice though that $f'(\frac{3}{2}^-) = -4\frac{3}{2} + 5 = -1 < \overline{f}'(u_L) < -\frac{1}{2} = f'(2)$, that is, $|\overline{f}'(u_L)| < 1$. Moreover, $\overline{f}'(u_R) = f'(u_R) = 1$. Hence, $\lambda_{\max}(1,2) = 1$ is a legitimate choice since $1 \ge \max(|\overline{f}'(u_L)|, |\overline{f}'(u_R)|)$.

Exercise 79.6 (Strong solutions). (i) Let u be a weak solution to (79.1) and assume that $u \in C^1(D\times[0,T^*))$. Let us first consider $\phi \in C_0^1(D\times(0,T^*))$. Integrating (79.7) by parts and applying the vanishing integral theorem in $D\times(0,T^*)$ (see Theorem 1.32), we infer that u solves (79.1). Let us now consider $\phi \in C_0^1(D\times[0,T^*))$. We infer that $\int_D \phi(\boldsymbol{x},0)(u(\boldsymbol{x},0) - u_0(\boldsymbol{x}))\,\mathrm{d}x = 0$. This implies that $\int_D \psi(\boldsymbol{x})(u(\boldsymbol{x},0) - u_0(\boldsymbol{x}))\,\mathrm{d}x = 0$ for all $\psi \in C_0^1(D)$ since one can always find $\phi \in C_0^1(D\times[0,T^*))$ such that $\phi(\boldsymbol{x},0) = \psi(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$. We conclude that $u(\boldsymbol{x},0) - u_0(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$ by invoking again the vanishing integral theorem, but this time in D. Hence, u solves (79.1). All the above operations are legitimate owing to the assumed smoothness of u.
(ii) Let $u \in C^1(D\times[0,T^*))$ be a strong solution to (79.1). Then (79.7) follows by multiplying (79.1) with ϕ and integrating by parts. All these operations are legitimate owing to the assumed smoothness of u.
(iii) Let $u \in C^1(D\times[0,T^*))$ be a strong solution to (79.1). Let $\psi \in C_0^1(D\times[0,T^*);\mathbb{R}_+)$ and let $(\eta, \boldsymbol{q})$ be an entropy pair. Assume that η is of class C^2. Let $\phi(\boldsymbol{x},t) = \psi(\boldsymbol{x},t)\eta'(u(\boldsymbol{x},t))$. Notice that $\phi \in C_0^1(D\times[0,T^*))$ because η is of class C^2. Multiplying (79.1) by ϕ and integrating by parts gives (79.10).

Exercise 79.7 (Method of characteristics). (i) Let us show that $u(x,t) = u_0(s(x,t))$ solves (79.1) for all $t \in [0,T^*)$. First $\partial_t u(x,t) = u_0'(s(x,t))\partial_t s(x,t)$ and $\partial_x f(u(x,t)) = f'(u(x,t))\partial_x u(x,t) = f'(u_0(s(x,t)))u_0'(s(x,t))\partial_x s(x,t)$. But the identity $x = f'(u_0(s))t + s$ implies that

$$\begin{aligned} 0 &= tf''(u_0(s))u_0'(s)\partial_t s + f'(u_0(s)) + \partial_t s,\\ 1 &= tf''(u_0(s))u_0'(s)\partial_x s + \partial_x s. \end{aligned}$$

Hence, we have

$$\begin{aligned} \partial_t s &= \frac{-f'(u_0(s))}{1 + f''(u_0(s))u_0'(s)t},\\ \partial_x s &= \frac{1}{1 + f''(u_0(s))u_0'(s)t}. \end{aligned}$$

In conclusion, we obtain

$$\begin{aligned} \partial_t u(x,t) + \partial_x(f(u(x,t))) &= u_0'(s(x,t))\frac{-f'(u_0(s))}{1 + f''(u_0(s))u_0'(s)t}\\ &\quad + f'(u_0(s(x,t)))u_0'(s(x,t))\frac{1}{1 + f''(u_0(s))u_0'(s)t}\\ &= 0. \end{aligned}$$

(ii) Along the segment $\{x = f'(u_0(s_0))t + s_0 \mid t \in [0, T^*]\}$, the function $s(x,t)$ solves

$$f'(u_0(s_0))t + s_0 - f'(u_0(s))t + s = 0.$$

But the unique solution to this equation is $s(x,t) = s_0$. Hence, $u(x,t) = u_0(s(x,t)) = u_0(s_0)$, so that $u(x,t)$ is constant along the straight segment in question.
(iii) Let us show that the solution found in Step (ii) is the entropy solution. Clearly $s \in C^1(D\times[0,T^*))$. Hence, $u(x,t) = u_0(s(x,t))$ is also in $C^1(D\times[0,T^*))$. This means that u is a strong solution to (79.1). We conclude by invoking Remark 79.13.

Exercise 79.8 (Shock interacting with an expansion wave). (i) The Riemann problem centered at $x = -1$ leads to a shock moving with speed $s = -\frac{1}{2}$. The Riemann problem centered at $x = 0$ leads to an expansion wave, and we have $u(x,t) = \frac{x}{t}$ in the sector $\{-1 \le \frac{x}{t} \le 0\}$. This construction is valid until the left boundary of this sector catches up with the shock. This happens at the time $t = 2$ at the position $x = -2$.
(ii) For all $t \ge 2$, let us describe the trajectory of the shock with the function $t \mapsto \chi(t)$, where $\chi(2) = -2$. Let us set

$$u_L(t) := \lim_{x\uparrow\chi(t)} u(x,t) = 0, \qquad u_R(t) := \lim_{x\downarrow\chi(t)} u(x,t) = \frac{\chi(t)}{t}.$$

Since the shock moves at the speed $\chi'(t)$, the Rankine–Hugoniot condition leads to

$$\chi'(t) = \frac{f(u_R(t)) - f(u_L(t))}{u_R(t) - u_L(t)} = \frac{\chi(t)}{2t}, \qquad \forall t \ge 2.$$

We infer that $\chi(t) = -(2t)^{\frac{1}{2}}$ for all $t \ge 2$, and the weak entropy solution is such that

$$u(x,t) = \begin{cases} 0 & \text{if } x \le \chi(t), \\ \frac{x}{t} & \text{if } \chi(t) < x \le 0, \\ 0 & \text{if } 0 \le x. \end{cases}$$

(iii) For all $t \in (0,2]$, we have

$$\int_{\mathbb{R}} u(x,t)\,\mathrm{d}x = \int_{-1-\frac{t}{2}}^{-t} -1\,\mathrm{d}x + \int_{-t}^{0} \frac{x}{t}\,\mathrm{d}x = -1 + \frac{t}{2} - \frac{1}{2}\frac{t^2}{t} = -1,$$

and for all $t \ge 2$ we have

$$\int_{\mathbb{R}} u(x,t)\,\mathrm{d}x = \int_{\chi(t)}^{0} \frac{x}{t}\,\mathrm{d}x = \frac{1}{2}\frac{\chi(t)^2}{t} = -1.$$

We have proved that $\int_{\mathbb{R}} u(x,t)\,\mathrm{d}x = \int_{\mathbb{R}} u_0(x)\,\mathrm{d}x$ for all $t \ge 0$.

Chapter 80

Hyperbolic systems

Exercises

Exercise 80.1 (1D linear system). (i) Let $u_0 \in L^\infty_{\rm loc}(\mathbb{R})$. Show that $u(x,t) := u_0(x-\lambda t)$ is a weak solution to the problem $\partial_t u + \lambda \partial_x u = 0$, $u(x,0) = u_0(x)$, i.e., $\int_0^\infty \int_{\mathbb{R}} u(\partial_t \phi + \lambda \partial_x \phi)\,\mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} u_0(x)\phi(x,0)\,\mathrm{d}x = 0$ for all $\phi \in C^1_0(\mathbb{R}\times\mathbb{R}_+)$. (ii) Let $\boldsymbol{u}_0 \in L^\infty_{\rm loc}(\mathbb{R};\mathbb{R}^m)$. Consider the one-dimensional linear system $\partial_t \boldsymbol{u} + \mathbb{A}\partial_x \boldsymbol{u} = 0$, $\boldsymbol{u}(x,0) = \boldsymbol{u}_0(x)$, $(x,t) \in \mathbb{R}\times\mathbb{R}_+$, where $\mathbb{A} \in \mathbb{R}^{m\times m}$ is diagonalizable in $\mathbb{R}$. Give a weak solution to this problem. (iii) Solve the 1D linear wave equation, i.e., consider $\mathbb{A} := \left(\begin{smallmatrix} 0 & 1 \\ c^2 & 0 \end{smallmatrix}\right)$.

Exercise 80.2 (Linear wave equation). Consider the matrix $\mathbb{A}(\boldsymbol{n}) := \left(\begin{smallmatrix} 0 & \boldsymbol{n}^\mathsf{T} \\ c^2\boldsymbol{n} & 0 \end{smallmatrix}\right)$, where $\boldsymbol{n}$ is a unit (column) vector in $\mathbb{R}^d$. Let $\{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{d-1}\}$ be such that $\{\boldsymbol{n},\boldsymbol{v}_1,\dots,\boldsymbol{v}_{d-1}\}$ is an orthonormal basis of $\mathbb{R}^d$. Show that $(c,(1,c\boldsymbol{n})^\mathsf{T})$, $(-c,(1,-c\boldsymbol{n})^\mathsf{T})$, $(0,(0,\boldsymbol{v}_1))$,..., $(0,(0,\boldsymbol{v}_{d-1}))$ are eigenpairs of $\mathbb{A}(\boldsymbol{n})$.

Exercise 80.3 (Entropy inequality). Let $\boldsymbol{u}_\epsilon$ be the smooth function satisfying $\partial_t \boldsymbol{u}_\epsilon + \nabla\cdot\mathbb{f}(\boldsymbol{u}_\epsilon) - \epsilon\Delta\boldsymbol{u}_\epsilon = 0$ in $D\times\mathbb{R}_+$, $\boldsymbol{u}_\epsilon(\cdot,0) = \boldsymbol{u}_0$ in D, with $\epsilon > 0$. Let $(\eta,\boldsymbol{q})$ be an entropy pair with $\eta \in C^2(\mathbb{R}^m;\mathbb{R})$. Prove that $\partial_t\eta(\boldsymbol{u}_\epsilon) + \nabla\cdot\boldsymbol{q}(\boldsymbol{u}_\epsilon) - \epsilon\Delta\eta(\boldsymbol{u}_\epsilon) \le 0$.

Exercise 80.4 (Convexity). Let $\sigma : \mathcal{T}\times\mathcal{E} \subset \mathbb{R}^2 \to \mathcal{S} \subset \mathbb{R}$ be a function of class C^2 such that $\partial_e\sigma(\tau,e) > 0$ for all $(\tau,e) \in \mathcal{T}\times\mathcal{E}$. (i) Show that there exists a function $\epsilon : \mathcal{T}\times\mathcal{S} \to \mathcal{E}$ such that $\sigma(\tau,\epsilon(\tau,s)) = s$ for all $(\tau,s) \in \mathcal{T}\times\mathcal{S}$ and ϵ is of class C^2. (ii) Show that $\epsilon(\tau,\sigma(\tau,e)) = e$ for all $(\tau,e) \in \mathcal{T}\times\mathcal{E}$. (iii) Show that the following statements are equivalent: (a) The function $\epsilon : \mathcal{T}\times\mathcal{S} \to \mathcal{E}$ is strictly convex; (b) The function $-\sigma : \mathcal{T}\times\mathcal{E} \to \mathcal{S}$ is strictly convex. (*Hint*: recall that a function $\phi : X \subset \mathbb{R}^m \to \mathbb{R}$ of class C^2 is convex in the open set X iff $D^2\phi(x)(h,h) > 0$ for all $h \in \mathbb{R}^m\backslash\{0\}$ and all $x \in X$.)

Exercise 80.5 (Euler). Recall from Example 80.10 the conserved variable $\boldsymbol{u} := (\rho,\boldsymbol{m}^\mathsf{T},E)^\mathsf{T}$, the specific internal energy $e(\boldsymbol{u}) := E/\rho - \frac{1}{2}\boldsymbol{m}^2/\rho^2$, and the function $\Phi(\boldsymbol{u}) := s(\rho,e(\boldsymbol{u}))$, where s is the specific entropy. (i) Is the function $\boldsymbol{u} \mapsto e(\boldsymbol{u})$ convex? (ii) Set $\Psi(\boldsymbol{u}) := -\rho\Phi(\boldsymbol{u})$. It is shown in Harten et al. [26, §3] that $\rho^{-1}K(D^2\Psi)K^\mathsf{T} = -C$, where $D^2\Psi$ is the Hessian matrix of Ψ and

$$K := \begin{pmatrix} 1 & \boldsymbol{v}^\mathsf{T} & \frac{1}{2}\boldsymbol{v}^2 + e \\ \boldsymbol{0} & \rho\mathbb{I}_d & \boldsymbol{m} \\ 0 & \boldsymbol{0}^\mathsf{T} & \rho \end{pmatrix}, \qquad C := \begin{pmatrix} \partial_{\rho\rho}s + \frac{2}{\rho}\partial_\rho s & \boldsymbol{0}^\mathsf{T} & \partial_{\rho e}s \\ \boldsymbol{0} & -\partial_e s\mathbb{I}_d & \boldsymbol{0} \\ \partial_{\rho e}s & \boldsymbol{0}^\mathsf{T} & \partial_{ee}s \end{pmatrix}.$$

Verify that K is invertible and C is negative definite. Show that the function $\boldsymbol{u} \mapsto \Psi(\boldsymbol{u})$ is strictly convex. (iii) Show that the set $B := \{\boldsymbol{u} \mid \rho > 0,\ e(\boldsymbol{u}) \geq 0\}$ is convex and that the set $B_r = \{\boldsymbol{u} \mid \rho > 0,\ e(\boldsymbol{u}) \geq 0,\ \Phi(\boldsymbol{u}) \geq r\}$ is convex for all $r \in \mathbb{R}$. (See also Exercise 83.3.) (iv) Let p be the pressure. Show that $\partial_\rho p(\rho, s) > 0$. (*Hint*: see Exercise 80.4 and recall that $\mathrm{d}\epsilon = T\,\mathrm{d}s - p\,\mathrm{d}\tau$.)

Exercise 80.6 (Wave equation blowup). Consider the linear wave equation in dimension three, $\partial_t u + \nabla\cdot\boldsymbol{v} = 0$, $\partial_t \boldsymbol{v} + \nabla u = \boldsymbol{0}$, with $u(\boldsymbol{x}, 0) = u_0(\|\boldsymbol{x}\|_{\ell^2})$, $\boldsymbol{v}(\boldsymbol{x}, 0) = \boldsymbol{0}$. Assume that $u_0 \in C^2(\mathbb{R}_+; \mathbb{R})$. (i) Show that u must solve $\partial_{tt} u - \nabla\cdot\nabla u = 0$. (ii) Let $f : \mathbb{R} \to \mathbb{R}$ be such that $f(s) := \frac{s}{2} u_0(s)$ if $s \geq 0$ and $f(s) := -f(-s)$ if $s \leq 0$. Let us write $r := \|\boldsymbol{x}\|_{\ell^2}$ and $\boldsymbol{e}_r := \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_{\ell^2}}$ if $\boldsymbol{x} \neq \boldsymbol{0}$. Show that $u(\boldsymbol{x}, t) = \frac{f(r+t)}{r} + \frac{f(r-t)}{r}$ and $\boldsymbol{v}(\boldsymbol{x}, t) = v(r, t)\boldsymbol{e}_r$, where the function $v(r, t) := -\frac{1}{r^2} \int_0^t \big(rf'(r+\tau) - f(r+\tau) + rf'(r-\tau) - f(r-\tau)\big)\,\mathrm{d}\tau$ solves the linear wave equation. (*Hint*: use spherical coordinates.) (iii) Compute $u(0, t)$ for $t > 0$. (iv) Let $\alpha \in (\frac{1}{2}, 1)$. Let $u_0(r) := 0$ if $0 \leq r \leq 1$, $u_0(r) := (r-1)^\alpha (2-r)^2$ if $r \in [1, 2]$, and $u_0(r) := 0$ if $2 \leq r$. Show that $u(\cdot, 1)$ is unbounded but $u(\cdot, 1) \in H^1(\mathbb{R}^3)$.

Exercise 80.7 (1D linear wave equation). Consider the 1D linear wave equation $\partial_t \boldsymbol{u} + \partial_x \mathbb{f}(\boldsymbol{u}) = 0$, where $\boldsymbol{u} := (\rho, v)^\mathsf{T}$, $\mathbb{f}(\boldsymbol{u}) := (\rho_0 v, p(\rho))^\mathsf{T}$, $p(\rho) := \frac{a^2}{\rho_0}\rho$, with the constants $\rho_0 > 0$ and $a > 0$. The purpose of the exercise is to show that the maximum principle does not hold true on ρ for the linear wave equation. (i) Show that the system is strictly hyperbolic. (ii) Are the characteristic families genuinely nonlinear or linearly degenerate? (iii) Consider the Riemann problem with $\boldsymbol{u}_L := (\rho_L, v_L)^\mathsf{T}$ and $\boldsymbol{u}_R := (\rho_R, v_R)^\mathsf{T}$. Express the two eigenvectors in terms of $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$. (iv) Solve the Riemann problem. (*Hint*: the solution is composed of three constant states separated by two contact discontinuities; apply the Rankine–Hugoniot condition two times.) (v) Give a condition on $v_L - v_R$ and $\rho_L - \rho_R$ so that $\min_{x\in\mathbb{R}} \rho(x, t) < \min(\rho_L, \rho_R)$. Give a condition on $v_L - v_R$ and $\rho_L - \rho_R$ so that $\min_{x\in\mathbb{R}} \rho(x, t) > \max(\rho_L, \rho_R)$. *Note*: this exercise shows that in general the maximum principle does not hold true on ρ for the linear wave equation.

Solution to exercises

Exercise 80.1 (1D linear system). (i) Let $\phi \in C_0^1(\mathbb{R}\times\mathbb{R}_+)$. Using the change of variable $x' = x - \lambda t$, we infer that

$$\begin{aligned}\int_0^\infty \int_{\mathbb{R}} u(\partial_t\phi + \lambda\partial_x\phi)\,\mathrm{d}x\mathrm{d}t &= \int_0^\infty \int_{\mathbb{R}} u_0(x - \lambda t)(\partial_t\phi + \lambda\partial_x\phi)\,\mathrm{d}x\mathrm{d}t \\ &= \int_0^\infty \int_{\mathbb{R}} u_0(x')(\partial_t\phi(x' + \lambda t, t) + \lambda\partial_x\phi(x' + \lambda t, t))\,\mathrm{d}x'\mathrm{d}t.\end{aligned}$$

Let $\psi(x', t) := \phi(x' + \lambda t, t)$. We have $\partial_t\psi(x', t) := \lambda\partial_x\phi(x' + \lambda t, t) + \partial_t\phi(x' + \lambda t, t)$. Using Fubini's theorem, we obtain

$$\begin{aligned}\int_0^\infty \int_{\mathbb{R}} u(\partial_t\phi + \lambda\partial_x\phi)\,\mathrm{d}x\mathrm{d}t &= \int_{\mathbb{R}} \int_0^\infty u_0(x')\partial_t\psi(x', t)\,\mathrm{d}x'\mathrm{d}t \\ &= \int_{\mathbb{R}} u_0(x') \int_0^\infty \partial_t\psi(x', t)\,\mathrm{d}x'\mathrm{d}t \\ &= -\int_{\mathbb{R}} u_0(x')\psi(x', 0)\,\mathrm{d}x' \\ &= -\int_{\mathbb{R}} u_0(x)\phi(x, 0)\,\mathrm{d}x.\end{aligned}$$

(ii) Let $\mathbb{A} = \mathbb{P}\mathbb{D}\mathbb{P}^{-1}$ be the spectral decomposition of $\mathbb{A}$. We have

$$(\partial_t \boldsymbol{u} + \mathbb{P}\mathbb{D}\mathbb{P}^{-1}\partial_x \boldsymbol{u} = \boldsymbol{0}) \iff (\partial_t \mathbb{P}^{-1}\boldsymbol{u} + \mathbb{D}\partial_x(\mathbb{P}^{-1}\boldsymbol{u}) = \boldsymbol{0}).$$

Setting $\boldsymbol{w} := \mathbb{P}^{-1}\boldsymbol{u}$, the above problem is thus equivalent to solving

$$\partial_t \boldsymbol{w} + \mathbb{D}\partial_x \boldsymbol{w} = \boldsymbol{0}, \qquad \boldsymbol{w}(x,0) = \boldsymbol{w}_0(x) := \mathbb{P}^{-1}\boldsymbol{u}_0(x).$$

Let $(\lambda_1, \dots, \lambda_m)$ be the eigenvalues of $\mathbb{A}$ and $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_m)$ be the associated eigenvectors, i.e., $\mathbb{P} = [\boldsymbol{v}_1 \dots \boldsymbol{v}_m]$. Let $(w_1(x,t), \dots, w_m(x,t))^{\mathsf{T}} = \boldsymbol{w}(x,t)$ and $(w_{01}(x), \dots, w_{0m}(x))^{\mathsf{T}} = \boldsymbol{w}_0(x)$. We obtain $\partial_t w_l + \lambda_l \partial_x w_l = 0$, $w_l(x,0) = w_{0l}(x)$ for all $l \in \{1\!:\!m\}$. Using Step (i), we infer that $w_l(x,t) = w_{0l}(x - \lambda_l t)$ is a weak solution to this problem. We conclude that

$$\boldsymbol{u}(x,t) = \sum_{l',l'' \in \{1:m\}} \mathbb{P}_{ll'}(\mathbb{P}^{-1})_{l'l''} u_{0l''}(x - \lambda_{l'} t).$$

(iii) The eigenpairs of $\mathbb{A}$ are $(c, (1,c)^{\mathsf{T}})$ and $(-c, (1-c)^{\mathsf{T}})$, i.e.,

$$\mathbb{P} = \begin{pmatrix} 1 & 1 \\ c & -c \end{pmatrix}, \quad \mathbb{P}^{-1} = \frac{1}{2c}\begin{pmatrix} c & 1 \\ c & -1 \end{pmatrix}.$$

Upon defining $\boldsymbol{u}_0(x) = (u_{01}(x), u_{02}(x))^{\mathsf{T}}$, we have

$$\boldsymbol{w}_0(x) = \mathbb{P}^{-1}\boldsymbol{u}_0(x) = \frac{1}{2c}\begin{pmatrix} cu_{01} + u_{02} \\ cu_{01} - u_{02} \end{pmatrix}.$$

Hence, we have

$$\boldsymbol{w}(x,t) = \frac{1}{2c}\begin{pmatrix} cu_{01}(x-ct) + u_{02}(x-ct) \\ cu_{01}(x+ct) - u_{02}(x+ct) \end{pmatrix},$$

and since $\boldsymbol{u}(x,t) = \mathbb{P}\boldsymbol{w}(x,t)$, we conclude that

$$\boldsymbol{u}(x,t) = \frac{1}{2c}\begin{pmatrix} cu_{01}(x-ct) + u_{02}(x-ct) + cu_{01}(x+ct) - u_{02}(x+ct) \\ c^2 u_{01}(x-ct) + cu_{02}(x-ct) - c^2 u_{01}(x+ct) + cu_{02}(x+ct) \end{pmatrix}.$$

Exercise 80.2 (Linear wave equation). We just verify the statement by doing the computation for each pair:

$$\mathbb{A}(\boldsymbol{n})\begin{pmatrix} 1 \\ c\boldsymbol{n} \end{pmatrix} = \begin{pmatrix} c \\ c^2\boldsymbol{n} \end{pmatrix} = c\begin{pmatrix} 1 \\ c\boldsymbol{n} \end{pmatrix},$$

$$\mathbb{A}(\boldsymbol{n})\begin{pmatrix} 1 \\ -c\boldsymbol{n} \end{pmatrix} = \begin{pmatrix} -c \\ c^2\boldsymbol{n} \end{pmatrix} = -c\begin{pmatrix} 1 \\ -c\boldsymbol{n} \end{pmatrix}.$$

Now, let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{d-1}$ be such that $(\boldsymbol{n}, \boldsymbol{v}_1, \dots, \boldsymbol{v}_{d-1})$ is an orthonormal basis of $\mathbb{R}^d$. We have

$$\mathbb{A}(\boldsymbol{n})\begin{pmatrix} 0 \\ \boldsymbol{v}_l \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_l \cdot \boldsymbol{n} \\ 0 \end{pmatrix} = 0\begin{pmatrix} 0 \\ \boldsymbol{v}_l \end{pmatrix}, \qquad \forall l \in \{1\!:\!d\!-\!1\}.$$

Exercise 80.3 (Entropy inequality). Let $(u_{i\epsilon})_{i \in \{1:m\}}$ be the components of $\boldsymbol{u}_\epsilon$. Let us multiply the equation $\partial_t \boldsymbol{u}_\epsilon + \nabla \cdot \mathbb{f}(\boldsymbol{u}_\epsilon) - \epsilon \Delta \boldsymbol{u}_\epsilon = 0$ by the column vector $\nabla \eta(\boldsymbol{u}_\epsilon)$ whose components are

$\partial_{v_i}\eta(\boldsymbol{u}_\epsilon)$ for all $i\in\{1{:}m\}$. Using that $\partial_{v_j}q_k(\boldsymbol{v})=\sum_{i\in\{1:m\}}\partial_{v_i}\eta(\boldsymbol{v})\partial_{v_j}(\mathbb{f}_{ik}(\boldsymbol{v}))$, we infer that

$$\begin{aligned}
0 &= \sum_{i\in\{1:m\}}\partial_{v_i}\eta(\boldsymbol{u}_\epsilon)\partial_t u_{i\epsilon}+\sum_{i\in\{1:m\}}\partial_{v_i}\eta(\boldsymbol{u}_\epsilon)\sum_{\substack{k\in\{1:d\}\\ j\in\{1:m\}}}\partial_{v_j}(\mathbb{f}_{ik}(\boldsymbol{u}_\epsilon))\partial_{x_k}u_{j\epsilon}\\
&\quad-\epsilon\sum_{i\in\{1:m\}}\partial_{v_i}\eta(\boldsymbol{u}_\epsilon)\Delta u_{i\epsilon}\\
&= \partial_t(\eta(\boldsymbol{u}_\epsilon))+\sum_{k\in\{1:d\}}\sum_{j\in\{1:m\}}\partial_{v_j}(q_k(\boldsymbol{u}_\epsilon))\partial_{x_k}u_{j\epsilon}\\
&\quad-\epsilon\sum_{i\in\{1:m\}}\nabla\cdot(\partial_{v_i}(\eta(\boldsymbol{u}_\epsilon))\nabla(u_{i\epsilon}))+\epsilon\sum_{i\in\{1:m\}}\nabla(\partial_{v_i}(\eta(\boldsymbol{u}_\epsilon))\cdot\nabla u_{i\epsilon}\\
&= \partial_t(\eta(\boldsymbol{u}_\epsilon))+\nabla\cdot(\boldsymbol{q}(\boldsymbol{u}_\epsilon))-\epsilon\Delta(\eta(\boldsymbol{u}_\epsilon))+\epsilon\sum_{k\in\{1:d\}}\sum_{i,j\in\{1:m\}}\partial_{v_iv_j}(\eta(\boldsymbol{u}_\epsilon))(\partial_{x_k}u_{j\epsilon})(\partial_{x_k}u_{i\epsilon}).
\end{aligned}$$

For every $k\in\{1{:}d\}$, the quantity $\sum_{i,j\in\{1:m\}}\partial_{v_iv_j}(\eta(\boldsymbol{u}_\epsilon))(\partial_{x_k}u_{j\epsilon})(\partial_{x_k}u_{i\epsilon})$ is nonnegative since η is convex, i.e., the matrix $\partial_{v_iv_j}(\eta(\boldsymbol{u}_\epsilon))$ is symmetric positive semidefinite. We have thus proved that

$$\partial_t(\eta(\boldsymbol{u}_\epsilon))+\nabla\cdot(\boldsymbol{q}(\boldsymbol{u}_\epsilon))-\epsilon\Delta(\eta(\boldsymbol{u}_\epsilon))\le 0.$$

Exercise 80.4 (Convexity). (i) Apply the implicit function theorem.
(ii) The definition of $\epsilon(\tau,s)$ implies that

$$\sigma(\tau,\epsilon(\tau,\sigma(\tau,e)))=\sigma(\tau,e),\qquad\forall(\tau,e)\in\mathcal{T}\times\mathcal{E}.$$

But σ being strictly monotone increasing with respect to the second variable, the above identity implies that $\epsilon(\tau,\sigma(\tau,e))=e$ for all $(\tau,e)\in\mathcal{T}\times\mathcal{E}$.
(iii) Let us first prove that (b) implies (a), i.e., we want to prove that the function ϵ is strictly convex if the function $-\sigma$ is strictly convex. Let $(\tau,s)\in\mathcal{T}\times\mathcal{S}$. Using the hint, we need to prove that $D^2\epsilon(\tau,s)\big((\tau',s'),(\tau',s')\big)>0$ for all $(\tau',s')\in\mathbb{R}^2\backslash\{(0,0)\}$. Using the Fréchet differential notation (see Appendix B) and applying the chain rule to the identity $\sigma(\tau,\epsilon(\tau,s))=s$, we obtain

$$D\sigma(\tau,\epsilon(\tau,s))\big(\tau',D\epsilon(\tau,s)(\tau',s')\big)=s'.$$

Applying the chain rule again, we obtain

$$\begin{aligned}
&D^2\sigma(\tau,\epsilon(\tau,s))\big((\tau',D\epsilon(\tau,s)(\tau',s')),(\tau',D\epsilon(\tau,s)(\tau',s'))\big)\\
&\qquad\qquad+D\sigma(\tau,\epsilon(\tau,s))\big(0,D^2\epsilon(\tau,s)((\tau',s'),(\tau',s'))\big)=0.
\end{aligned}$$

Using that $D\sigma(\tau,e)(0,h)=\partial_e\sigma(\tau,e)h$ and $\partial_e\sigma(\tau,e)>0$, we infer that

$$D^2\epsilon(\tau,s)((\tau',s'),(\tau',s'))=-\frac{D^2\sigma(\tau,e)\big((\tau',e'),(\tau',e')\big)}{\partial_e\sigma(\tau,e)},$$

with $e:=\epsilon(\tau,s)$ and $e':=D\epsilon(\tau,s)(\tau',s')$. Since $-\sigma$ is strictly convex by assumption, the above identity proves that ϵ is strictly convex. Let us now prove the converse statement. Let $(\tau,e)\in\mathcal{T}\times\mathcal{E}$ and let $(\tau',e')\in\mathbb{R}^2$. Let us set $s:=\sigma(\tau,e)$ and $s':=D\sigma(\tau,e)(\tau',e')$. Reasoning as above, we obtain

$$D^2\sigma(\tau,e)\big((\tau',e'),(\tau',e')\big)=-\partial_e\sigma(\tau,e)D^2\epsilon(\tau,s)\big((\tau',s'),(\tau',s')\big).$$

This shows that if ϵ is strictly convex, $-\sigma$ is strictly convex.

Exercise 80.5 (Euler). (i) The function $e(\rho, \boldsymbol{m}, E)$ is clearly not convex since $\partial_{m_i m_i} e(\rho, \boldsymbol{m}, E) = -1/\rho^2 < 0$ for all $i \in \{1{:}d\}$.
(ii) The matrix K is upper triangular, and its diagonal entries are nonzero since $\rho > 0$. Hence K is invertible. The matrix C is negative definite iff the following matrix is negative definite

$$C' := \begin{pmatrix} \partial_{\rho\rho} s + \frac{2}{\rho}\partial_\rho s & \partial_{\rho e} s \\ \partial_{\rho e} s & \partial_{ee} s \end{pmatrix}.$$

Setting $\sigma(\tau, e) := s(\tau^{-1}, e)$, we obtain

$$C' := \begin{pmatrix} \tau^4 \partial_{\tau\tau}\sigma & -\tau^2 \partial_{\tau e}\sigma \\ -\tau^2 \partial_{\tau e}\sigma & \partial_{ee}\sigma \end{pmatrix}.$$

We then infer that C' is negative definite since the function $\sigma(\tau, e)$ is strictly concave. Hence, C is negative definite. It is now clear that $D^2\Psi$ is positive definite, so that the function Ψ is strictly convex.
(iii) Let $\theta \in [0,1]$ and $\boldsymbol{u}_0 := (\rho_0, \boldsymbol{m}_0^\mathsf{T}, E_0)^\mathsf{T}$ and $\boldsymbol{u}_1 := (\rho_1, \boldsymbol{m}_1^\mathsf{T}, E_1)^\mathsf{T}$ be two members of the set $\mathcal{B} := \{\boldsymbol{u} \mid \rho > 0,\ e(\boldsymbol{u}) \geq 0\}$. Let us set $\boldsymbol{u}_\theta := \theta\boldsymbol{u}_0 + (1-\theta)\boldsymbol{u}_1 = (\rho_\theta, \boldsymbol{m}_\theta^\mathsf{T}, E_\theta)^\mathsf{T}$. Clearly, we have $\rho_\theta = \theta\rho_0 + (1-\theta)\rho_1 > 0$. Moreover, we have

$$\begin{aligned}
(\theta\rho_0 + (1-\theta)\rho_1)(\theta E_0 + (1-\theta)E_1) &= \theta^2\rho_0 E_0 + \theta(1-\theta)(\rho_0 E_1 + \rho_1 E_0) + (1-\theta)^2\rho_1 E_1 \\
&\geq \theta^2 \frac{1}{2}\boldsymbol{m}_0^2 + \theta(1-\theta)\Big(\frac{\rho_0}{\rho_1}\frac{1}{2}\boldsymbol{m}_1^2 + \frac{\rho_1}{\rho_0}\frac{1}{2}\boldsymbol{m}_0^2\Big) + (1-\theta)^2\frac{1}{2}\boldsymbol{m}_1^2 \\
&\geq \theta^2 \frac{1}{2}\boldsymbol{m}_0^2 + \theta(1-\theta)\boldsymbol{m}_0{\cdot}\boldsymbol{m}_1 + (1-\theta)^2\frac{1}{2}\boldsymbol{m}_1^2 \\
&\quad + \theta(1-\theta)\left(\frac{\rho_0}{\rho_1}\frac{1}{2}\boldsymbol{m}_1^2 + \frac{\rho_1}{\rho_0}\frac{1}{2}\boldsymbol{m}_0^2 - \boldsymbol{m}_0{\cdot}\boldsymbol{m}_1\right).
\end{aligned}$$

Using that $\boldsymbol{m}_0{\cdot}\boldsymbol{m}_1 \leq \frac{1}{2}(\frac{\rho_1}{\rho_0}\boldsymbol{m}_0^2 + \frac{\rho_0}{\rho_1}\boldsymbol{m}_1^2)$, we infer that

$$(\theta\rho_0 + (1-\theta)\rho_1)(\theta E_0 + (1-\theta)E_1) \geq \frac{1}{2}(\theta\boldsymbol{m}_0 + (1-\theta)\boldsymbol{m}_1)^2,$$

which shows that $\rho_\theta E_\theta \geq \frac{1}{2}\boldsymbol{m}_\theta^2$. Since $\rho_\theta > 0$, we conclude that $e(\boldsymbol{u}_\theta) = E_\theta/\rho_\theta - \frac{1}{2}\boldsymbol{m}_\theta^2/\rho_\theta^2 \geq 0$. We have thus proved that $\boldsymbol{u}_\theta \in \mathcal{B}$.

Another way to prove the above result consists of observing that $\boldsymbol{u} \mapsto \rho e = \rho(E/\rho - \frac{1}{2}\boldsymbol{m}^2/\rho^2)$ is a concave function because $D^2(\rho e)(\boldsymbol{v}, \boldsymbol{v}) = -\frac{1}{\rho}(\frac{\boldsymbol{m}}{\rho}a - \boldsymbol{b})^2$ for all $\boldsymbol{v} := (a, \boldsymbol{b}^\mathsf{T}, c)^\mathsf{T} \in \mathbb{R}{\times}\mathbb{R}^d{\times}\mathbb{R}$. Let $\theta \in [0,1]$ and $\boldsymbol{u}_0 := (\rho_0, \boldsymbol{m}_0^\mathsf{T}, E_0)^\mathsf{T}$, $\boldsymbol{u}_1 := (\rho_1, \boldsymbol{m}_1^\mathsf{T}, E_1)^\mathsf{T}$ be two members of the set $\mathcal{B}$. We have

$$(\rho e)(\theta\boldsymbol{u}_0 + (1-\theta)\boldsymbol{u}_1) \geq \theta(\rho e)(\boldsymbol{u}_0) + (1-\theta)(\rho e)(\boldsymbol{u}_1) \geq 0,$$

which proves that $e(\theta\boldsymbol{u}_0 + (1-\theta)\boldsymbol{u}_1) \geq 0$ because $\rho(\theta\boldsymbol{u}_0 + (1-\theta)\boldsymbol{u}_1) > 0$.
To establish the convexity of the set $\mathcal{B}_r$, we proceed as above. We first observe that the function $\boldsymbol{u} \mapsto \rho\Phi(\boldsymbol{u}) - \rho r$ is concave. Let $\theta \in [0,1]$ and $\boldsymbol{u}_0 = (\rho_0, \boldsymbol{m}_0^\mathsf{T}, E_0)^\mathsf{T}$, $\boldsymbol{u}_1 = (\rho_1, \boldsymbol{m}_1^\mathsf{T}, E_1)^\mathsf{T}$ be two members of the set $\mathcal{B}_r$. We have

$$\begin{aligned}
(\rho\Phi - \rho r)(\theta\boldsymbol{u}_0 + (1-\theta)\boldsymbol{u}_1) &\geq \theta\rho\Phi(\boldsymbol{u}_0) + (1-\theta)\rho\Phi(\boldsymbol{u}_1) - (\theta\rho_0 + (1-\theta)\rho_1)r \\
&\geq \theta\rho_0 r + (1-\theta)\rho_1 r - (\theta\rho_0 + (1-\theta)\rho_1)r \\
&\geq 0,
\end{aligned}$$

where we used that $\rho\Phi(\boldsymbol{u}_0) \geq \rho_0 r$ and $\rho\Phi(\boldsymbol{u}_1) \geq \rho_1 r$. This proves that $\Phi(\theta\boldsymbol{u}_0 + (1-\theta)\boldsymbol{u}_1) \geq r$ because $\rho(\theta\boldsymbol{u}_0 + (1-\theta)\boldsymbol{u}_1) > 0$.
(iv) Using the hint, we have $p(\tau, s) = -\partial_\tau \epsilon(\tau, s)$. With an obvious abuse of notation, we obtain $\partial_\rho p(\tau(\rho), s) = -\frac{1}{\rho^2}\partial_\tau p(\tau, s) = \frac{1}{\rho^2}\partial_{\tau\tau}\epsilon(\tau, s)$, and we know that $\partial_{\tau\tau}\epsilon(\tau, s) > 0$ since ϵ is strictly convex. Hence, $\partial_\rho p(\tau, s) > 0$ for all $(\tau, s) \in (0, \infty)\times\mathbb{R}$.

Exercise 80.6 (Wave equation blowup). (i) One can eliminate $\boldsymbol{v}$ by taking the time derivative of the first equation, taking the divergence of the second equation, and taking the difference between the results. This yields $\partial_{tt} u - \nabla\cdot\nabla u = 0$.
(ii) Let us compute $\partial_t \boldsymbol{v} = \partial_t v\, \boldsymbol{e}_r$ and $\nabla u = \partial_r u\, \boldsymbol{e}_r$. Using the chain rule, we have

$$\partial_t v = -\frac{1}{r^2}\big(rf'(r+t) - f(r+t) + rf'(r-t) - f(r-t)\big),$$
$$\partial_r u = \frac{1}{r^2}\big(rf'(r+t) - f(r+t) + rf'(r-t) - f(r-t)\big).$$

Hence, $\partial_t v + \partial_r u = 0$, i.e., $\partial_t \boldsymbol{v} + \nabla u = \boldsymbol{0}$. Let us compute $\partial_t u$ and $\nabla\cdot\boldsymbol{v} = \frac{1}{r^2}\partial_r(r^2 v)$. We obtain

$$\begin{aligned}
\partial_t u &= \frac{1}{r}\big(f'(r+t) - f'(r-t)\big),\\
\frac{1}{r^2}\partial_r(r^2 v) &= -\frac{1}{r^2}\int_0^t \big(rf''(r+\tau) + rf''(r-\tau)\big)\,\mathrm{d}\tau\\
&= -\frac{1}{r^2}\Big[rf'(r+\tau) - rf'(r-\tau)\Big]_{\tau=0}^{\tau=t}\\
&= -\frac{1}{r^2}\big(rf'(r+t) - rf'(r-t)\big).
\end{aligned}$$

Hence, $\partial_t u + \frac{1}{r^2}\partial_r(r^2 v) = 0$, so that $\partial_t u + \nabla\cdot\boldsymbol{v} = 0$.
(iii) Using that f is odd, we have $f(r-t) = -f(t-r)$ and

$$u(0,t) = \lim_{r\downarrow 0}\Big(\frac{f(r+t)}{r} + \frac{f(r-t)}{r}\Big) = \lim_{r\downarrow 0}\frac{f(t+r) - f(t-r)}{r} = 2f'(t).$$

Recalling that $f(t) = \frac{t}{2}u_0(t)$ for all $t > 0$, we obtain $u(0,t) = tu_0'(t) + u_0(t)$ for all $t > 0$.
(iv) We obtain for $r = \|\boldsymbol{x}\|_{\ell^2} \leq 1$,

$$2u(\boldsymbol{x}, 1) = \frac{(r+1)u_0(r+1)}{r} - \frac{(1-r)u_0(1-r)}{r} = (r+1)r^{\alpha-1}(1-r)^2,$$

since $u_0(1+r) = r^\alpha(1-r)^2$ and $u_0(1-r) = 0$. Hence, $\lim_{r\downarrow 0} u(\boldsymbol{x}, 1) = \infty$ because $\alpha < 1$. The function $u(\boldsymbol{x}, 1)$ is also nonzero for $r \in [2, 3]$ where we have $2u(\boldsymbol{x}, 1) = r^{-1}(r-1)(r-2)^\alpha(3-r)^2$. Therefore, we have

$$\nabla u(\boldsymbol{x}, 1) = (r^{\alpha-2} + \mathcal{O}(r^{\alpha-1}))\boldsymbol{e}_r,$$

for $r = \|\boldsymbol{x}\|_{\ell^2} \leq 3$ and $\nabla u(\boldsymbol{x}, 1) = \boldsymbol{0}$ for $r \geq 3$. Hence, we obtain

$$|u|^2_{H^1(\mathbb{R}^3)} = \int_0^1 r^{2(\alpha-2)+2}\,\mathrm{d}r + \int_0^1 r^{2(\alpha-1)+2}\,\mathrm{d}r + c.$$

The quantity $|u|^2_{H^1(\mathbb{R}^3)}$ is bounded since $\alpha > \frac{1}{2}$, i.e., $2(\alpha-2)+2 > -1$.

Exercise 80.7 (1D linear wave equation). (i) We have $d = 1$, $\boldsymbol{n} = \boldsymbol{e}_x$, and

$$\mathbb{A}(\boldsymbol{n}) = \begin{pmatrix} 0 & \rho_0 \\ \frac{a^2}{\rho_0} & 0 \end{pmatrix}.$$

There are two distinct eigenvalues $\lambda_1 = -a$, $\lambda_2 = a$ with corresponding eigenvectors $\boldsymbol{r}_1 = (\rho_0, -a)^\mathsf{T}$ and $\boldsymbol{r}_2 = (\rho_0, a)^\mathsf{T}$. The eigenvalues being distinct, $\mathbb{A}(\boldsymbol{n})$ is diagonalizable. This proves that the system is strictly hyperbolic.
(ii) The two characteristic families are linearly degenerate since $D\lambda_1(\boldsymbol{u}) = 0$ and $D\lambda_2(\boldsymbol{u}) = 0$ for all $\boldsymbol{u} \in \mathbb{R}^2$.
(iii) Let us express $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$ in terms of the two eigenvectors. We have $\boldsymbol{u}_L = \alpha_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2$ with

$$\alpha_1 = \frac{a\rho_L - \rho_0 v_L}{2a\rho_0}, \qquad \alpha_2 = \frac{a\rho_L + \rho_0 v_L}{2a\rho_0}.$$

Similarly, $\boldsymbol{u}_R = \beta_1 \boldsymbol{r}_1 + \beta_2 \boldsymbol{r}_2$ with

$$\beta_1 = \frac{a\rho_R - \rho_0 v_R}{2a\rho_0}, \qquad \beta_2 = \frac{a\rho_R + \rho_0 v_R}{2a\rho_0}.$$

(iv) Since the two characteristic families are linearly degenerate, the solution to the Riemann problem is composed of three constant states separated by two contact discontinuities moving at speed $-a$ and a. Let $\boldsymbol{u}^*$ be the middle state. The Rankine–Hugoniot condition implies that

$$\begin{aligned} \mathbb{f}(\boldsymbol{u}_L) - \mathbb{f}(\boldsymbol{u}^*) &= \mathbb{A}(\boldsymbol{n})(\boldsymbol{u}_L - \boldsymbol{u}^*) = -a(\boldsymbol{u}_L - \boldsymbol{u}^*), \\ \mathbb{f}(\boldsymbol{u}^*) - \mathbb{f}(\boldsymbol{u}_R) &= \mathbb{A}(\boldsymbol{n})(\boldsymbol{u}^* - \boldsymbol{u}_R) = a(\boldsymbol{u}^* - \boldsymbol{u}_R). \end{aligned}$$

This, in turn, implies that there are $\mu_1, \mu_2 \in \mathbb{R}$ such that $\boldsymbol{u}_L - \boldsymbol{u}^* = \mu_1 \boldsymbol{r}_1$ and $\boldsymbol{u}^* - \boldsymbol{u}_R = \mu_2 \boldsymbol{r}_2$. Hence, we have

$$\mu_1 \boldsymbol{r}_1 + \mu_2 \boldsymbol{r}_2 = \boldsymbol{u}_L - \boldsymbol{u}_R = \alpha_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2 - \beta_1 \boldsymbol{r}_1 - \beta_2 \boldsymbol{r}_2.$$

Since the two eigenvectors are linearly independent, this implies that

$$\mu_1 = \alpha_1 - \beta_1, \qquad \mu_2 = \alpha_2 - \beta_2.$$

Thus, we have

$$\boldsymbol{u}^* = \beta_1 \boldsymbol{r}_1 + \beta_2 \boldsymbol{r}_2 + (\alpha_2 - \beta_2)\boldsymbol{r}_2 = \beta_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2.$$

In conclusion, we have

$$\begin{aligned} \rho^* &= \frac{a\rho_R - \rho_0 v_R}{2a} + \frac{a\rho_L + \rho_0 v_L}{2a} = \frac{a(\rho_R + \rho_L) + \rho_0(v_L - v_R)}{2a}, \\ v^* &= -\frac{a\rho_R - \rho_0 v_R}{2\rho_0} + \frac{a\rho_L + \rho_0 v_L}{2\rho_0} = \frac{a(\rho_L - \rho_R) + \rho_0(v_L + v_R)}{2\rho_0}. \end{aligned}$$

The solution is given by

$$\boldsymbol{u}(x,t) = \begin{cases} \boldsymbol{u}_L & \text{if } x \le -at, \\ \boldsymbol{u}^* & \text{if } -at < x \le at, \\ \boldsymbol{u}_R & \text{if } at < x. \end{cases}$$

(v) We have $\rho^* < \min(\rho_L, \rho_R)$ if

$$\min(\rho_L, \rho_R) > \frac{a(\rho_R + \rho_L) + \rho_0(v_L - v_R)}{2a} = \frac{\rho_R + \rho_L}{2} + \frac{\rho_0(v_L - v_R)}{2a},$$

which can be rewritten as

$$\frac{\rho_0(v_L - v_R)}{2a} < \min(\rho_L, \rho_R) - \frac{\rho_R + \rho_L}{2} = -\frac{1}{2}|\rho_L - \rho_R|.$$

Similarly, we have $\rho^* > \max(\rho_L, \rho_R)$ if

$$\max(\rho_L, \rho_R) > \frac{a(\rho_R + \rho_L) + \rho_0(v_L - v_R)}{2a} = \frac{\rho_R + \rho_L}{2} + \frac{\rho_0(v_L - v_R)}{2a},$$

which can be rewritten as

$$\frac{\rho_0(v_L - v_R)}{2a} > \max(\rho_L, \rho_R) - \frac{\rho_R + \rho_L}{2} = \frac{1}{2}|\rho_L - \rho_R|.$$

In conclusion, we have $\rho^* < \min(\rho_L, \rho_R)$ if $\rho_0(v_L - v_R) < -a|\rho_L - \rho_R|$ and $\rho^* > \max(\rho_L, \rho_R)$ if $\rho_0(v_L - v_R) > a|\rho_L - \rho_R|$.

Chapter 81

First-order approximation

Exercises

Exercise 81.1 (1D approximation). Consider the one-dimensional problem $\partial_t u + \nabla\cdot\boldsymbol{f}(u) = 0$ with $D := (-1,1)$ and $\boldsymbol{f}(v) := f(v)\boldsymbol{e}_x$. Let $I \in \mathbb{N}$, $I \geq 3$, and consider the mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$ for all $i \in \{1{:}I{-}1\}$, such that $-1 =: x_1 < \cdots < x_I := 1$, with $h_i := x_{i+1} - x_i$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions on $\mathcal{T}_h$. (i) Compute $\boldsymbol{c}_{i,i-1}$ and $\boldsymbol{n}_{i,i-1}$ for all $i \in \{2{:}I\}$, $\boldsymbol{c}_{i,i}$ and m_i for all $i \in \{2{:}I{-}1\}$, and $\boldsymbol{c}_{i,i+1}$ and $\boldsymbol{n}_{i,i+1}$ for all $i \in \{1{:}I{-}1\}$. (ii) Assuming that f is convex, compute $\lambda_{\max}(\boldsymbol{n}_{i,i-1}, \mathsf{U}_i^n, \mathsf{U}_{i-1}^n)$, $\lambda_{\max}(\boldsymbol{n}_{i-1,i}, \mathsf{U}_{i-1}^n, \mathsf{U}_i^n)$, $\lambda_{\max}(\boldsymbol{n}_{i,i+1}, \mathsf{U}_i^n, \mathsf{U}_{i+1}^n)$, and $\lambda_{\max}(\boldsymbol{n}_{i+1,i}, \mathsf{U}_{i+1}^n, \mathsf{U}_i^n)$. (iii) Compute $d_{i,i-1}^n$ and $d_{i,i+1}^n$. (iv) Justify (81.11).

Exercise 81.2 (Symmetry). Let $i \in \mathcal{A}_h^\circ$. (i) Show that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ for all $j \in \mathcal{I}(i)$. (ii) Show that $\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2} = \lambda_{\max}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}$.

Exercise 81.3 (Average matrix). Let $\mathcal{A} \subset \mathbb{R}^m$ and $\mathbb{f} \in \mathrm{Lip}(\mathcal{A}; \mathbb{R}^{m\times d})$ with components $(\mathbb{f}_{kl})_{k\in\{1:m\},l\in\{1:d\}}$. Let $\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathbb{R}^m$ and consider the matrix $\mathbb{A}_{kk'} := \int_0^1 \partial_{v_{k'}}(\mathbb{f}\cdot\boldsymbol{n})_k(\boldsymbol{u}_R + \theta(\boldsymbol{u}_L - \boldsymbol{u}_R))\,\mathrm{d}\theta$. (i) Show that $(\mathbb{f}(\boldsymbol{u}_L) - \mathbb{f}(\boldsymbol{u}_R))\cdot\boldsymbol{n} = \mathbb{A}(\boldsymbol{u}_L - \boldsymbol{u}_R)$. (ii) Assume from now on that $m := 1$ and set $A := \mathbb{A}$, i.e., we are working with scalar equations. Compute A if $u_L \neq u_R$, $\lim_{u_R\to u_L} A$ and $\lim_{u_L\to u_R} A$ assuming that $\mathbb{f}$ is C^1. (iii) Under which conditions do we have $|A| = \lambda_{\max}(\boldsymbol{n}, u_L, u_R)$ if $\mathbb{f}$ is either convex or concave? (*Hint*: see §79.2.) (iv) Take $d_{ij}^n := |A|$ in (81.9) with $\boldsymbol{n} := \boldsymbol{n}_{ij}$, $u_L := \mathsf{U}_i^n$, and $u_R := \mathsf{U}_j^n$. Prove that Theorem 81.8 still holds true if τ is small enough.

Exercise 81.4 (Entropy glitch). Consider the one-dimensional problem $\partial_t u + \nabla\cdot(f(u)\boldsymbol{e}_x) = 0$ with $D := (-1,1)$ and data $u_0(x) := -1$ if $x \leq 0$ and $u_0(x) := 1$ otherwise. Let $I \in \mathbb{N}\backslash\{0\}$ be an even number, and consider the mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$, $i \in \{1{:}I{-}1\}$, such that $-1 =: x_1 < \cdots < x_I := 1$ and $x_{\frac{I}{2}} \leq 0 < x_{\frac{I}{2}+1}$. Let $h_i := x_{i+1} - x_i$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions on $\mathcal{T}_h$. (i) Take $d_{ij}^n := \|\boldsymbol{c}_{ij}\|_{\ell^2}|(f(\mathsf{U}_i^n) - f(\mathsf{U}_j^n))/(\mathsf{U}_i^n - \mathsf{U}_j^n)|$ if $\mathsf{U}_i^n \neq \mathsf{U}_j^n$ and $d_{ij}^n := \|\boldsymbol{c}_{ij}\|_{\ell^2}|f'(\mathsf{U}_i^n)|$ otherwise. Prove that Theorem 81.8 still holds true if τ is small enough. (ii) Consider Burgers' flux $\boldsymbol{f}(u) := \frac{1}{2}u^2\boldsymbol{e}_x$. Take $u_h^0(x) := \sum_{i\in\mathcal{A}_h} \mathsf{U}_i^0\varphi_i(x)$ with $\mathsf{U}_i^0 := -1$ if $i \leq \frac{1}{2}I$ and $\mathsf{U}_i^0 := 1$ if $i \geq \frac{1}{2}I + 1$. Using the above definition of d_{ij}^n, show that the scheme (81.9) gives $u_h^n = u_h^0$ for any $n \geq 0$. Comment on this result.

Solution to exercises

Exercise 81.1 (1D approximation). (i) We have $\boldsymbol{c}_{i,i-1} = -\frac{1}{2}\boldsymbol{e}_x$ and $\boldsymbol{n}_{i,i-1} = -\boldsymbol{e}_x$ for all $i \in \{2{:}I\}$, $\boldsymbol{c}_{i,i} = \boldsymbol{0}$ and $m_i = \frac{h_{i-1}+h_i}{2}$ for all $i \in \{2{:}I{-}1\}$, and $\boldsymbol{c}_{i,i+1} = \frac{1}{2}\boldsymbol{e}_x$ and $\boldsymbol{n}_{i,i+1} = \boldsymbol{e}_x$ for all $i \in \{1{:}I{-}1\}$.
(ii) Since the function $f = \boldsymbol{e}_x{\cdot}\boldsymbol{f}$ is convex, we have

$$\lambda_{\max}(\boldsymbol{e}_x, u_L, u_R) = \begin{cases} \left|\frac{f(u_L)-f(u_R)}{u_L-u_R}\right| & \text{if } u_L > u_R, \\ \max(|f'(u_L)|, |f'(u_R)|) & \text{otherwise.} \end{cases}$$

Since $\boldsymbol{n}_{i-1,i} = \boldsymbol{e}_x$, we obtain

$$\lambda_{\max}(\boldsymbol{n}_{i-1,i}, \mathsf{U}_{i-1}^n, \mathsf{U}_i^n) = \begin{cases} \left|\frac{f(\mathsf{U}_{i-1}^n)-f(\mathsf{U}_i^n)}{\mathsf{U}_{i-1}^n-\mathsf{U}_i^n}\right| & \text{if } \mathsf{U}_{i-1}^n > \mathsf{U}_i^n, \\ \max(|f'(\mathsf{U}_{i-1}^n)|, |f'(\mathsf{U}_i^n)|) & \text{otherwise.} \end{cases}$$

Since $\boldsymbol{n}_{i,i-1} = -\boldsymbol{e}_x$ so that $\boldsymbol{n}_{i,i-1}{\cdot}\boldsymbol{f} = -f$ is a concave function, we have

$$\lambda_{\max}(\boldsymbol{n}_{i,i-1}, \mathsf{U}_i^n, \mathsf{U}_{i-1}^n) = \begin{cases} \left|\frac{f(\mathsf{U}_{i-1}^n)-f(\mathsf{U}_i^n)}{\mathsf{U}_{i-1}^n-\mathsf{U}_i^n}\right| & \text{if } \mathsf{U}_{i-1}^n > \mathsf{U}_i^n, \\ \max(|f'(\mathsf{U}_{i-1}^n)|, |f'(\mathsf{U}_i^n)|) & \text{otherwise.} \end{cases}$$

We observe that $\lambda_{\max}(-\boldsymbol{e}_x, \mathsf{U}_{i-1}^n, \mathsf{U}_i^n) = \lambda_{\max}(\boldsymbol{e}_x, \mathsf{U}_i^n, \mathsf{U}_{i-1}^n)$. Using the above relations with the index i shifted by 1, we finally infer that

$$\lambda_{\max}(\boldsymbol{n}_{i,i+1}, \mathsf{U}_i^n, \mathsf{U}_{i+1}^n) = \begin{cases} \left|\frac{f(\mathsf{U}_{i+1}^n)-f(\mathsf{U}_i^n)}{\mathsf{U}_{i+1}^n-\mathsf{U}_i^n}\right| & \text{if } \mathsf{U}_i^n > \mathsf{U}_{i+1}^n, \\ \max(|f'(\mathsf{U}_{i+1}^n)|, |f'(\mathsf{U}_i^n)|) & \text{otherwise,} \end{cases}$$

and

$$\lambda_{\max}(\boldsymbol{n}_{i+1,i}, \mathsf{U}_{i+1}^n, \mathsf{U}_i^n) = \begin{cases} \left|\frac{f(\mathsf{U}_i^n)-f(\mathsf{U}_{i+1}^n)}{\mathsf{U}_i^n-\mathsf{U}_{i+1}^n}\right| & \text{if } \mathsf{U}_i^n > \mathsf{U}_{i+1}^n, \\ \max(|f'(\mathsf{U}_i^n)|, |f'(\mathsf{U}_{i+1}^n)|) & \text{otherwise.} \end{cases}$$

(iii) The above computations show that

$$d_{i,i-1}^n = \frac{1}{2}\begin{cases} \left|\frac{f(\mathsf{U}_{i-1}^n)-f(\mathsf{U}_i^n)}{\mathsf{U}_{i-1}^n-\mathsf{U}_i^n}\right| & \text{if } \mathsf{U}_{i-1}^n > \mathsf{U}_i^n, \\ \max(|f'(\mathsf{U}_{i-1}^n)|, |f'(\mathsf{U}_i^n)|) & \text{otherwise,} \end{cases}$$

$$d_{i,i+1}^n = \frac{1}{2}\begin{cases} \left|\frac{f(\mathsf{U}_{i+1}^n)-f(\mathsf{U}_i^n)}{\mathsf{U}_{i+1}^n-\mathsf{U}_i^n}\right| & \text{if } \mathsf{U}_i^n > \mathsf{U}_{i+1}^n, \\ \max(|f'(\mathsf{U}_{i+1}^n)|, |f'(\mathsf{U}_i^n)|) & \text{otherwise.} \end{cases}$$

(iv) In the case of the linear transport equation, we have $f(u) = \beta u$, so that $f'(u) = \beta$. Hence, $d_{i,i-1}^n = d_{i,i+1}^n = |\beta|$. Using that

$$\mathsf{U}_i^{n+1} = \mathsf{U}_i^n + \frac{\tau}{2m_i}(f(\mathsf{U}_{i-1}^n) - f(\mathsf{U}_{i+1}^n)) + \frac{\tau}{m_i}d_{i,i-1}^n(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n) + \frac{\tau}{m_i}d_{i,i+1}^n(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n),$$

we obtain

$$\begin{aligned} \mathsf{U}_i^{n+1} &= \mathsf{U}_i^n + \frac{\tau}{2m_i}\beta(\mathsf{U}_{i-1}^n - \mathsf{U}_{i+1}^n) + \frac{\tau}{2m_i}|\beta|(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n) + \frac{\tau}{2m_i}|\beta|(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n) \\ &= \mathsf{U}_i^n + \frac{\tau}{2m_i}(\beta + |\beta|)(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n) + \frac{\tau}{2m_i}(|\beta| - \beta)(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n). \end{aligned}$$

This is (81.11).

Exercise 81.2 (Symmetry). (i) Since $i \in \mathcal{A}_h^\circ$, we have $\varphi_{i|\partial D} = 0$, so that integrating by parts, we infer that

$$\boldsymbol{c}_{ij} = \int_D \varphi_i \nabla\varphi_j \,\mathrm{d}x = -\int_D \varphi_j \nabla\varphi_i \,\mathrm{d}x = -\boldsymbol{c}_{ji}.$$

(ii) The identity $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ implies that $\boldsymbol{n}_{ij} = -\boldsymbol{n}_{ji}$. This shows that up to the change of variable $x \to -x$, the Riemann problem with flux $\mathbb{f}\cdot\boldsymbol{n}_{ij}$ and data $(\boldsymbol{u}_L, \boldsymbol{u}_R)$ has the same solution as the Riemann problem with flux $\mathbb{f}\cdot\boldsymbol{n}_{ji}$ and data $(\boldsymbol{u}_R, \boldsymbol{u}_L)$. This implies that the maximum wave speeds in the two Riemann problems are identical.

Exercise 81.3 (Average matrix). (i) Let us define $\psi(\theta) := \mathbb{f}(\boldsymbol{u}_R + \theta(\boldsymbol{u}_L - \boldsymbol{u}_R))\cdot\boldsymbol{n}$. Let $(\mathbb{f}\cdot\boldsymbol{n})_k$, for all $k \in \{1\!:\!m\}$, be the components of $\mathbb{f}\cdot\boldsymbol{n}$ and let $(\boldsymbol{u}_L - \boldsymbol{u}_R)_{k'}$, for all $k' \in \{1\!:\!m\}$, be the components of $\boldsymbol{u}_L - \boldsymbol{u}_R$. Using the chain rule, we have

$$\partial_\theta\psi_k(\theta) = \sum_{k'\in\{1:m\}} \partial_{v_{k'}}(\mathbb{f}\cdot\boldsymbol{n})_k(\boldsymbol{u}_R + \theta(\boldsymbol{u}_L - \boldsymbol{u}_R))(\boldsymbol{u}_L - \boldsymbol{u}_R)_{k'}.$$

This proves that

$$\begin{aligned}(\mathbb{f}(\boldsymbol{u}_L) - \mathbb{f}(\boldsymbol{u}_R))\cdot\boldsymbol{n} &= \int_0^1 \partial_\theta\psi_k(\theta)\,\mathrm{d}\theta\\ &= \sum_{k'\in\{1:m\}} \left(\int_0^1 \partial_{v_{k'}}(\mathbb{f}\cdot\boldsymbol{n})_k(\boldsymbol{u}_R + \theta(\boldsymbol{u}_L - \boldsymbol{u}_R))\,\mathrm{d}\theta\right)(\boldsymbol{u}_L - \boldsymbol{u}_R)_{k'}\\ &= \mathbb{A}(\boldsymbol{u}_L - \boldsymbol{u}_R).\end{aligned}$$

(ii) Let us take $m := 1$ from now on. Then $A := \mathbb{A}$ is a scalar. From Step (i), we infer that $A = \frac{(\boldsymbol{f}(u_L) - \boldsymbol{f}(u_R))\cdot\boldsymbol{n}}{u_L - u_R}$ if $u_L \neq u_R$. Moreover, $\lim_{u_L\to u_R} A = (\boldsymbol{f}\cdot\boldsymbol{n})'(u_R)$ and $\lim_{u_R\to u_L} A = (\boldsymbol{f}\cdot\boldsymbol{n})'(u_L)$.
(iii) Let $\underbar{$f$}$ and $\bar{f}$ be the lower and upper convex envelopes of $\boldsymbol{f}\cdot\boldsymbol{n}$ over the interval $\mathrm{conv}(u_L, u_R)$, respectively. Notice first that $A = \underbar{$f$}'$ if $\boldsymbol{f}(v)\cdot\boldsymbol{n}$ is concave and $A = \bar{f}'$ if $\boldsymbol{f}(v)\cdot\boldsymbol{n}$ is convex. In conclusion, if $\boldsymbol{f}(v)\cdot\boldsymbol{n}$ is convex, we have $|A| = \lambda_{\max}(\boldsymbol{n}, u_L, u_R)$ only if $u_L > u_R$, and, if $\boldsymbol{f}(v)\cdot\boldsymbol{n}$ is concave, we have $|A| = \lambda_{\max}(\boldsymbol{n}, u_L, u_R)$ only if $u_L < u_R$. The solution to the Riemann problem is a shock in both cases.
(iv) Let us set $A_{ij} := \frac{\boldsymbol{f}(\mathsf{U}_j)\cdot\boldsymbol{n}_{ij} - \boldsymbol{f}(\mathsf{U}_i)\cdot\boldsymbol{n}_{ij}}{\mathsf{U}_j - \mathsf{U}_i}$ if $\mathsf{U}_j \neq \mathsf{U}_i$ and $A_{ij} := \boldsymbol{f}'(\mathsf{U}_i)\cdot\boldsymbol{n}_{ij}$ otherwise. Let $d_{ij}^n := \max(|A_{ij}|\|\boldsymbol{c}_{ij}\|_{\ell^2}, |A_{ji}|\|\boldsymbol{c}_{ji}\|_{\ell^2})$ in (81.9). We obtain

$$\begin{aligned}m_i\frac{\mathsf{U}_i^{n+1} - \mathsf{U}_i^n}{\tau} &= \sum_{j\in\mathcal{I}(i)} \Big(-\boldsymbol{f}(\mathsf{U}_j^n)\cdot\boldsymbol{c}_{ij} + d_{ij}^n\mathsf{U}_j^n\Big)\\ &= \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \Big((\boldsymbol{f}(\mathsf{U}_i^n) - \boldsymbol{f}(\mathsf{U}_j^n))\cdot\boldsymbol{n}_{ij}\|\boldsymbol{c}_{ij}\|_{\ell^2} + d_{ij}^n(\mathsf{U}_j^n - \mathsf{U}_i^n)\Big)\\ &= \sum_{j\in\mathcal{I}(i)\setminus\{i\}} (d_{ij}^n - A_{ij}\|\boldsymbol{c}_{ij}\|_{\ell^2})(\mathsf{U}_j^n - \mathsf{U}_i^n).\end{aligned}$$

This, in turn, implies that

$$\mathsf{U}_i^{n+1} = \left(1 - \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{\tau}{m_i}(d_{ij}^n - A_{ij}\|\boldsymbol{c}_{ij}\|_{\ell^2})\right)\mathsf{U}_i^n + \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{\tau}{m_i}(d_{ij}^n - A_{ij}\|\boldsymbol{c}_{ij}\|_{\ell^2})\mathsf{U}_j^n.$$

Provided τ is small enough so that $1 - \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{\tau}{m_i}(d_{ij}^n - A_{ij}\|\boldsymbol{c}_{ij}\|_{\ell^2}) \geq 0$, we have a convex combination, because $d_{ij}^n - A_{ij}\|\boldsymbol{c}_{ij}\|_{\ell^2} \geq 0$ by definition of d_{ij}^n. This means that U_i^{n+1} is in the convex hull of $\{\mathsf{U}_j^n\}_{j\in\mathcal{I}(i)}$. In conclusion, the local maximum principle holds true.

Exercise 81.4 (Entropy glitch). (i) The statement is proved in the solution of Exercise 81.3. (ii) Let us consider the approximate initial data $u_h^0 := \sum_{i\in\mathcal{A}_h} \mathsf{U}_i^0\varphi_i(x)$ with $\mathsf{U}_i^0 := -1$ if $i \leq \frac{1}{2}I$ and $\mathsf{U}_i^0 := 1$ if $i \geq \frac{1}{2}I + 1$. Let $n \geq 0$. The definition for the update U_i^{n+1} is

$$\mathsf{U}_i^{n+1} = \mathsf{U}_i^n + \frac{\tau}{2m_i}(f(\mathsf{U}_{i-1}^n) - f(\mathsf{U}_{i+1}^n)) + \frac{\tau}{m_i}d_{ii-1}^n(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n) + \frac{\tau}{m_i}d_{ii+1}^n(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n),$$

for all $i \in \{2{:}I{-}1\}$, whereas $\mathsf{U}_1^n = -1$ and $\mathsf{U}_I^n = 1$. It is clear that $\mathsf{U}_i^1 = \mathsf{U}_i^0$ for all $i \leq \frac{I}{2} - 1$ and all $\frac{I}{2} + 2 \leq i$. For $i = \frac{I}{2}$, we have

$$\mathsf{U}_{\frac{I}{2}-1}^0 = -1, \quad \mathsf{U}_{\frac{I}{2}}^0 = -1, \quad \mathsf{U}_{\frac{I}{2}+1}^0 = 1,$$

giving

$$\begin{aligned} f(\mathsf{U}_{\frac{I}{2}+1}^0) - f(\mathsf{U}_{\frac{I}{2}-1}^0) &= \frac{1}{2}(1-1) = 0, \\ d_{\frac{I}{2},\frac{I}{2}-1}^0 = \frac{1}{2}|f'(\mathsf{U}_{\frac{I}{2}}^0)| = \frac{1}{2}, &\quad d_{\frac{I}{2},\frac{I}{2}+1}^0 = 0. \end{aligned}$$

Hence, $\mathsf{U}_{\frac{I}{2}}^1 = \mathsf{U}_{\frac{I}{2}}^0$. Similarly, for $i = \frac{I}{2} + 1$, we have

$$\mathsf{U}_{\frac{I}{2}}^0 = -1, \quad \mathsf{U}_{\frac{I}{2}+1}^0 = 1, \quad \mathsf{U}_{\frac{I}{2}+2}^0 = 1,$$

giving

$$\begin{aligned} f(\mathsf{U}_{\frac{I}{2}+2}^0) - f(\mathsf{U}_{\frac{I}{2}}^0) &= \frac{1}{2}(1-1) = 0, \\ d_{\frac{I}{2}+1,\frac{I}{2}}^0 = 0, &\quad d_{\frac{I}{2}+1,\frac{I}{2}+2}^0 = \frac{1}{2}|f'(\mathsf{U}_{\frac{I}{2}+1}^0)| = \frac{1}{2}. \end{aligned}$$

Hence, $\mathsf{U}_{\frac{I}{2}+1}^1 = \mathsf{U}_{\frac{I}{2}+1}^0$. In conclusion, $u_h^1 = u_h^0$, so that $u_h^n = u_h^0$, for all $n \geq 0$. This shows that the numerical solution is a stationary discontinuity, whereas it should be an approximation of an expansion wave. Hence, the method does not converge to the entropy solution.

Chapter 82

Higher-order approximation

Exercises

Exercise 82.1 ($(\alpha-\beta)$ vs. Butcher representation). (i) Consider the ERK scheme defined by the Butcher tableau (82.11), i.e., the matrix $\mathcal{A} \in \mathbb{R}^{s\times s}$ and the vector $b \in \mathbb{R}^s$. Consider the matrix $\mathbb{A} := \begin{pmatrix} \mathcal{A} & \mathbf{0} \\ b^\mathsf{T} & 0 \end{pmatrix}$ of order $(s+1)$, with $\mathbf{0} := (0,\dots,0)^\mathsf{T} \in \mathbb{R}^s$. Set $u^{(i)} := u^n + \tau \sum_{j\in\{1:i-1\}} a_{ij}k_j$ for all $i \in \{1{:}s\}$, where a_{ij} are the entries of the matrix $\mathcal{A}$. Consider the vectors $\mathsf{U} := (u^{(1)},\dots,u^{(s)},u^{n+1})^\mathsf{T}$ and $\mathsf{F}(\mathsf{U}) := (L(t_n+c_1\tau,u^{(1)}),\dots,L(t_n+c_s\tau,u^{(s)}),0)^\mathsf{T}$. Show that $\mathsf{U} = u^n\mathsf{E} + \tau\mathbb{A}\mathsf{F}(\mathsf{U})$ with $\mathsf{E} := (1,\dots,1)^\mathsf{T} \in \mathbb{R}^{s+1}$. (ii) Consider the scheme defined by the $(\alpha\text{-}\beta)$ representation (82.6) with $\gamma_k := c_{k+1}$ for all $k \in \{0{:}s{-}1\}$. Let $\mathbb{a}$ and $\mathbb{b}$ be the $(s+1)\times(s+1)$ strictly lower triangular matrices with entries $\mathbb{a}_{i+1,k+1} := \alpha_{ik}$, $\mathbb{b}_{i+1,k+1} := \beta_{ik}$ for all $1 \le k+1 \le i \le s$. Show that $(\mathbb{I}-\mathbb{a})\mathsf{E} = \mathsf{E}_1$ with $\mathsf{E}_1 := (1,0,\dots,0)^\mathsf{T} \in \mathbb{R}^{s+1}$. (iii) Consider the vectors $\mathsf{W} := (w^{(0)},\dots,w^{(s)})^\mathsf{T}$, $\mathsf{F}(\mathsf{W}) := (L(t_n+c_1\tau,w^{(0)}),\dots,L(t_n+c_s\tau,w^{(s-1)}),0)^\mathsf{T}$. Show that $\mathsf{W} = u^n\mathsf{E} + \tau(\mathbb{I}-\mathbb{a})^{-1}\mathbb{b}\mathsf{F}(\mathsf{W})$. (iv) Compute the matrices $\mathbb{a}$, $\mathbb{b}$, and $(\mathbb{I}-\mathbb{a})^{-1}\mathbb{b}$ for the SSPRK(2,2) scheme. *Note*: this exercise shows that given the $(\alpha\text{-}\beta)$ representation (82.6), there is only one associated Butcher tableau. But given a Butcher tableau, there may be more than one $(\alpha\text{-}\beta)$ representation since the factorization $\mathbb{A} = (\mathbb{I}-\mathbb{a})^{-1}\mathbb{b}$ may be nonunique.

Exercise 82.2 (Quadratic approximation). (i) Give the expression of the reference shape functions for the Lagrange element $(\widehat{K},\mathbb{P}_{2,1},\{\widehat{\sigma}_1,\widehat{\sigma}_2,\widehat{\sigma}_3\})$ where $\widehat{K} := [0,1]$, $\widehat{\sigma}_1(\widehat{p}) := \widehat{p}(0)$, $\widehat{\sigma}_2(\widehat{p}) := \widehat{p}(\frac{1}{2})$, $\widehat{\sigma}_3(\widehat{p}) := \widehat{p}(1)$. (ii) Compute the reference mass matrix $M_{\widehat{K}}$ with entries $\int_{\widehat{K}} \widehat{\theta}_i(\widehat{x})\widehat{\theta}_j(\widehat{x})\,\mathrm{d}\widehat{x}$. (iii) Compute the lumped reference mass matrix $\overline{M}_{\widehat{K}}$. What should be the sum of the entries of $\overline{M}_{\widehat{K}}$? (iv) Let $D := (0,1)$. Let $N_\mathrm{e} \ge 1$, $I := 2N_\mathrm{e}+1$, and let $0 =: x_1 < \ldots < x_I := 1$. Consider the mesh $\mathcal{T}_h$ composed of the cells $K_m := [x_{2m-1},x_{2m+1}]$, $\forall m \in \{1{:}N_\mathrm{e}\}$. Let $h_m := x_{2m+1} - x_{2m-1}$. Let $P_2^\mathrm{g}(\mathcal{T}_h)$ be the H^1-conforming space based on $\mathcal{T}_h$ using quadratic polynomials. Give the expression of the global shape functions of $P_2^\mathrm{g}(\mathcal{T}_h)$ associated with the Lagrange nodes $\{x_i\}_{i\in\mathcal{A}_h}$ with $\mathcal{A}_h := \{1{:}I\}$. (v) Give the coefficients of the consistent mass matrix. (vi) Give the coefficients of the lumped mass matrix. What should be the sum of the entries of M^L? (vii) Is it possible to use the above Lagrange basis together with the theory described in §81.1.2 to approximate hyperbolic systems? (viii) Is it possible to apply Corollary 81.9 and Corollary 81.15?

Exercise 82.3 (Quadratic Bernstein approximation). Consider the following reference shape

functions on $\widehat{K} := [0,1]$:

$$\widehat{\theta}_1(\widehat{x}) := (1-\widehat{x})^2, \quad \widehat{\theta}_2(\widehat{x}) := 2\widehat{x}(1-\widehat{x}), \quad \widehat{\theta}_3(\widehat{x}) := \widehat{x}^2.$$

(i) Show that $\{\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3\}$ is a basis of $\mathbb{P}_{2,1}$. Show that these functions satisfy the partition of unity property and that $\widehat{p}(\widehat{x}) = \widehat{p}(0)\widehat{\theta}_1(\widehat{x}) + \widehat{p}(\frac{1}{2})\widehat{\theta}_2(\widehat{x}) + \widehat{p}(1)\widehat{\theta}_3(\widehat{x})$ for all $\widehat{p} \in \mathbb{P}_{1,1}$. (ii)-(viii) Redo Questions (ii)-(viii) of Exercise 82.2 with the above reference shape functions.

Exercise 82.4 (Gap estimates). The objective is to prove Lemma 82.10. (i) Let $\mathsf{U}^{\mathrm{L},n+1}$ be the update given by (81.9) with the low-order graph viscosity d_{ij}^{L}. Consider the auxiliary states $\overline{\mathsf{U}}_{ij}^n := \frac{1}{2}(\mathsf{U}_j^n + \mathsf{U}_i^n) - (\boldsymbol{f}(\mathsf{U}_j^n) - \boldsymbol{f}(\mathsf{U}_i^n))\cdot\frac{\boldsymbol{c}_{ij}}{2d_{ij}^{\mathrm{L},n}}$ defined in the proof of Theorem 81.8 for all $j \in \mathcal{I}(i)$, and set $\mathsf{U}_i^{*,n} := \frac{1}{\gamma_i^n}\sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\overline{\mathsf{U}}_{ij}^n$. Show that

$$\mathsf{U}_i^{n+1} = (1-\gamma_i^n)\mathsf{U}_i^n + \gamma_i^n\mathsf{U}_i^{*,n} + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}}(d_{ij}^n - d_{ij}^{\mathrm{L},n})(\mathsf{U}_j^n - \mathsf{U}_i^n).$$

(ii) Using that $\mathsf{U}_{ij}^{*,n} \le \mathsf{U}_i^{\mathrm{M},n}$, $d_{ij}^n \le d_{ij}^{\mathrm{L},n}$, and $\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0$, show that

$$\mathsf{U}_i^{n+1} \le \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})\Big((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i^-)}(d_{ij}^{\mathrm{L},n} - d_{ij}^n)\Big).$$

(iii) Using that $d_{ij}^n \ge d_{ij}^{\mathrm{L},n}\psi_i^n$ and $\psi_i^n \ge 0$, prove the upper bound in (82.23). (iv) Prove the lower bound in (82.23).

Solution to exercises

Exercise 82.1 ($(\alpha-\beta)$ vs. Butcher representation). (i) Using the notation from (82.12), we set

$$\begin{aligned} u^{(1)} &:= u^n \\ u^{(i)} &:= u^n + \tau\sum_{j\in\{1:i-1\}} a_{ij}k_j, \qquad \forall i \in \{2:s\}. \end{aligned}$$

Then we have $k_i = L(t_n + c_i\tau, u^{(i)})$ for all $i \in \{1:s\}$. This implies that

$$\begin{aligned} u^{(i)} &= u^n + \sum_{j\in\{1:i-1\}} a_{ij}L(t_n + c_j\tau, u^{(j)}), \qquad \forall i \in \{1:s\}, \\ u^{n+1} &= u^n + \sum_{i\in\{1:s\}} \tau b_i L(t_n + c_i\tau, u^{(i)}). \end{aligned}$$

Let us define the vectors

$$\begin{aligned} \mathsf{U} &:= (u^{(1)}, \dots, u^{(s)}, u^{n+1})^\mathsf{T}, \\ \mathsf{F}(\mathsf{U}) &:= (L(t_n + c_1\tau, u^{(1)}), \dots, L(t_n + c_s\tau, u^{(s)}), 0)^\mathsf{T}. \end{aligned}$$

Setting $\mathsf{E} := (1,\dots,1)^\mathsf{T} \in \mathbb{R}^{s+1}$, the above identities can be rewritten as

$$\mathsf{U} = u^n\mathsf{E} + \tau\mathbb{A}\mathsf{F}(\mathsf{U}).$$

(ii) The identity $(\mathbb{I}-\mathbb{a})\mathsf{E} = \mathsf{E}_1$ is a consequence of $\sum_{j\in\{1:i-1\}}\alpha_{ij} = 1$ for all $i \in \{1{:}s\}$.
(iii) Let $\mathsf{W} := (w^{(0)},\dots,w^{(s)})^\mathsf{T}$, $\mathsf{F}(\mathsf{W}) := (L(t_n{+}c_1\tau, w^{(0)}),\dots,L(t_n{+}c_s\tau,w^{(s-1)}),0)^\mathsf{T}$. The scheme (82.6) is equivalent to

$$\mathsf{W} = u^n\mathsf{E}_1 + \mathbb{a}\mathsf{W} + \tau\mathbb{b}\mathsf{F}(\mathsf{W}).$$

The identity $(\mathbb{I}-\mathbb{a})\mathsf{E} = \mathsf{E}_1$, in turn, implies that $\mathsf{W} = u^n\mathsf{E} + \tau(\mathbb{I}-\mathbb{a})^{-1}\mathbb{b}\mathsf{F}(\mathsf{W})$.
(iv) For the SSPRK(2,2) method, we have

$$\mathbb{a} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ \frac12 & \frac12 & 0 \end{pmatrix}, \qquad \mathbb{b} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \frac12 & 0 \end{pmatrix},$$

leading to

$$(\mathbb{I}-\mathbb{a})^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & \frac12 & 1 \end{pmatrix}, \qquad (\mathbb{I}-\mathbb{a})^{-1}\mathbb{b} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ \frac12 & \frac12 & 0 \end{pmatrix}.$$

We recover the leftmost Butcher tableau in (82.13).

Exercise 82.2 (Quadratic approximation). (i) We have

$$\widehat\theta_1(\widehat x) = (1-\widehat x)(1-2\widehat x), \quad \widehat\theta_2(\widehat x) = 4\widehat x(1-\widehat x), \quad \widehat\theta_3(\widehat x) = \widehat x(2\widehat x - 1).$$

(ii) We have

$$M_{\widehat K} = \frac{1}{15}\begin{pmatrix} 2 & 1 & -\frac12 \\ 1 & 8 & 1 \\ -\frac12 & 1 & 2 \end{pmatrix}.$$

(iii) Recall that the entries of $\overline{M}_{\widehat K}$ are $\overline{m}_{\widehat K,i}\delta_{ij}$, where $\overline{m}_{\widehat K,i} := \sum_{j\in\mathcal{N}} m_{\widehat K,ij}$ and $\mathcal{N} := \{1,2,3\}$. This yields

$$\overline{M}_{\widehat K} = \frac16\begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Owing to the partition of unity, we have $\overline{m}_{\widehat K,i} = \sum_{j\in\mathcal{N}}\int_{\widehat K}\widehat\theta_i\widehat\theta_j\,\mathrm{d}\widehat x = \int_{\widehat K}\widehat\theta_i\,\mathrm{d}\widehat x$. Hence, the sum of the entries of $\overline{M}_{\widehat K}$ is $\sum_{i\in\mathcal{N}}\overline{m}_{\widehat K,i} = \sum_{i\in\mathcal{N}}\int_{\widehat K}\widehat\theta_i\,\mathrm{d}\widehat x = |\widehat K| = 1$. This is indeed the result that we have obtained above.
(iv) Setting $\widehat x := \frac{x-x_{2m-1}}{h_m}$, we have

$$\varphi_{2m-1|K_m}(x) = \widehat\theta_1(\widehat x), \quad \varphi_{2m|K_m}(x) = \widehat\theta_2(\widehat x), \quad \varphi_{2m+1|K_m}(x) = \widehat\theta_3(\widehat x).$$

(v) Let m_{ij} denote the generic coefficient of the mass matrix. We have

$$m_{11} = \frac{2}{15}h_1, \quad m_{12} = \frac{1}{15}h_1, \quad m_{13} = -\frac{1}{30}h_1.$$

For all $m \in \{1{:}N_\mathrm{e}\}$, we have

$$m_{2m,2m-1} = \frac{1}{15}h_m, \quad m_{2m,2m} = \frac{8}{15}h_m, \quad m_{2m,2m+1} = \frac{1}{15}h_m.$$

If $N_{\mathrm{e}} \geq 2$, then we have for all $m \in \{2\!:\!N_{\mathrm{e}}\}$,

$$m_{2m-1,2m-3} = -\frac{1}{30}h_{m-1}, \quad m_{2m-1,2m-2} = \frac{1}{15}h_{m-1},$$
$$m_{2m-1,2m-1} = \frac{2}{15}(h_{m-1}+h_m),$$
$$m_{2m-1,2m} = \frac{1}{15}h_m, \quad m_{2m-1,2m+1} = -\frac{1}{30}h_m.$$

Finally, we have

$$m_{2N_{\mathrm{e}}+1,2N_{\mathrm{e}}-1} = -\frac{1}{30}h_{N_{\mathrm{e}}}, \quad m_{2N_{\mathrm{e}}+1,2N_{\mathrm{e}}} = \frac{1}{15}h_{N_{\mathrm{e}}}, \quad m_{2N_{\mathrm{e}}+1,2N_{\mathrm{e}}+1} = \frac{2}{15}h_{N_{\mathrm{e}}}.$$

(vi) Let $\overline{m}_i$ denote the generic diagonal coefficient of the lumped mass matrix. We have

$$\overline{m}_1 = \frac{1}{6}h_1.$$

For all $m \in \{1\!:\!N_{\mathrm{e}}\}$, we have

$$\overline{m}_{2m} = \frac{2}{3}h_m.$$

If $N_{\mathrm{e}} \geq 2$, then we have for all $m \in \{2\!:\!N_{\mathrm{e}}\}$,

$$\overline{m}_{2m-1} = \frac{1}{6}(h_{m-1}+h_m).$$

Finally, we have

$$\overline{m}_{2N_{\mathrm{e}}+1} = \frac{1}{6}h_{N_{\mathrm{e}}}.$$

Using the partition of unity, the sum of all the entries of the lumped mass matrix $\overline{M}$ is equal to $\sum_{i\in\mathcal{A}_h}\sum_{j\in\mathcal{A}_h}\int_D \varphi_i\varphi_j\,\mathrm{d}x = \sum_{i\in\mathcal{A}_h}\int_D\varphi_i\,\mathrm{d}x = |D| = 1$. This is indeed what we have obtained since

$$\begin{aligned}\sum_{i\in\mathcal{A}_h}\overline{m}_i &= \frac{1}{6}h_1 + \frac{1}{6}h_{N_{\mathrm{e}}} + \sum_{m\in\{2:N_{\mathrm{e}}\}}\frac{1}{6}(h_{m-1}+h_m) + \sum_{m\in\{1:N_{\mathrm{e}}\}}\frac{2}{3}h_m\\ &= \Big(\frac{2}{3}+\frac{1}{3}\Big)\sum_{m\in\{1:N_{\mathrm{e}}\}}h_m = |D|.\end{aligned}$$

(vi) It is possible to use the above Lagrange basis together with the theory described in §81.1.2 to approximate hyperbolic systems because the coefficients of the lumped mass matrix are positive (see (81.5)). This is the only required condition.
(viii) It is not possible to apply Corollary 81.9 and Corollary 81.15 because the shape functions can take negative values.

Exercise 82.3 (Quadratic Bernstein approximation). (i) Assume that there are a_1, a_2, a_3 s.t. $a_1\widehat{\theta}_1(\widehat{x}) + a_2\widehat{\theta}_2(\widehat{x}) + a_3\widehat{\theta}_3(\widehat{x}) = 0$ for all $\widehat{x}\in\widehat{K}$. Then

$$a_3\widehat{x}^2 - 2a_2\widehat{x}^2 + a_1\widehat{x}^2 + 2a_2\widehat{x} - 2a_1\widehat{x} + a_1 = 0,$$

i.e., $a_3 - 2a_2 + a_1 = 0$, $2a_2 - 2a_1 = 0$, and $a_1 = 0$. This immediately implies that $a_1 = 0$, $a_2 = 0$, and $a_3 = 0$. Hence, the functions $\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3$ are linearly independent. Finally, we verify that

$$\widehat{\theta}_1(\widehat{x}) + \widehat{\theta}_2(\widehat{x}) + \widehat{\theta}_3(\widehat{x}) = (\widehat{x}+1-\widehat{x})^2 = 1, \qquad \forall\widehat{x}\in\widehat{K},$$

thereby proving the partition of unity property. Let now $\widehat{p} \in \mathbb{P}_{2,1}$. There are a_1, a_2, a_3 s.t. $\widehat{p} = a_1\widehat{\theta}_1(\widehat{x}) + a_2\widehat{\theta}_2(\widehat{x}) + a_3\widehat{\theta}_3(\widehat{x})$. Then $\widehat{p}(0) = a_1$, $\widehat{p}(\frac{1}{2}) = \frac{1}{4}a_1 + \frac{1}{2}a_2 + \frac{1}{4}a_3$, and $\widehat{p}(1) = a_3$. Assume now that $\widehat{p} \in \mathbb{P}_{1,1}$. Then $\widehat{p}(\frac{1}{2}) = \frac{1}{2}(\widehat{p}(0) + \widehat{p}(1)) = \frac{1}{2}(a_1 + a_3)$, i.e., $\frac{1}{4}a_1 + \frac{1}{2}a_2 + \frac{1}{4}a_3 = \frac{1}{2}(a_1 + a_3)$. This implies that $a_2 = \frac{1}{2}(a_1 + a_3) = \widehat{p}(\frac{1}{2})$. Hence, for all $\widehat{p} \in \mathbb{P}_{1,1}$, we have

$$\widehat{p}(\widehat{x}) = \widehat{p}(0)\widehat{\theta}_1(\widehat{x}) + \widehat{p}(\tfrac{1}{2})\widehat{\theta}_2(\widehat{x}) + \widehat{p}(1)\widehat{\theta}_3(\widehat{x}).$$

(ii) We have

$$M_{\widehat{K}} = \frac{1}{5}\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{2} & \frac{2}{3} & \frac{1}{2} \\ \frac{1}{6} & \frac{1}{2} & 1 \end{pmatrix}.$$

(iii) We have

$$\overline{M}_{\widehat{K}} = \frac{1}{3}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Clearly, $\sum_{i\in\mathcal{N}} \overline{M}_{\widehat{K},ii} = |\widehat{K}| = 1$ as expected.
(iv) Setting $\widehat{x} := \frac{x - x_{2m-1}}{h_m}$, we have

$$\varphi_{2m-1|K_m}(x) = \widehat{\theta}_1(\widehat{x}), \quad \varphi_{2m|K_m}(x) = \widehat{\theta}_2(\widehat{x}), \quad \varphi_{2m+1|K_m}(x) = \widehat{\theta}_3(\widehat{x}).$$

(v) Let m_{ij} denote the generic coefficient of the mass matrix. We have

$$m_{11} = \frac{1}{5}h_1, \quad m_{12} = \frac{1}{10}h_1, \quad m_{13} = \frac{1}{30}h_1.$$

For all $m \in \{1\!:\!N_{\mathrm{e}}\}$, we have

$$m_{2m,2m-1} = \frac{1}{10}h_m, \quad m_{2m,2m} = \frac{2}{15}h_m, \quad m_{2m,2m+1} = \frac{1}{10}h_m.$$

If $N_{\mathrm{e}} \geq 2$, then we have for all $m \in \{2\!:\!N_{\mathrm{e}}\}$,

$$\begin{aligned} m_{2m-1,2m-3} &= \frac{1}{30}h_{m-1}, \quad m_{2m-1,2m-2} = \frac{1}{10}h_{m-1}, \\ m_{2m-1,2m-1} &= \frac{1}{5}(h_{m-1} + h_m), \\ m_{2m-1,2m} &= \frac{1}{10}h_m, \quad m_{2m-1,2m+1} = \frac{1}{30}h_m. \end{aligned}$$

Finally, we have

$$m_{2N_{\mathrm{e}}+1,2N_{\mathrm{e}}-1} = \frac{1}{30}h_{N_{\mathrm{e}}}, \quad m_{2N_{\mathrm{e}}+1,2N_{\mathrm{e}}} = \frac{1}{10}h_{N_{\mathrm{e}}}, \quad m_{2N_{\mathrm{e}}+1,2N_{\mathrm{e}}+1} = \frac{1}{5}h_{N_{\mathrm{e}}}.$$

(vi) Let $\overline{m}_i$ denote the generic diagonal coefficient of the lumped mass matrix. We have

$$\overline{m}_1 = \frac{1}{3}h_1.$$

For all $m \in \{1\!:\!N_{\mathrm{e}}\}$, we have

$$\overline{m}_{2m} = \frac{1}{3}h_m.$$

If $N_{\mathrm{e}} \geq 2$, then we have for all $m \in \{2\!:\!N_{\mathrm{e}}\}$,

$$\overline{m}_{2m-1} = \frac{1}{3}(h_{m-1} + h_m).$$

Finally, we have

$$\overline{m}_{2N_{\mathrm{e}}+1} = \frac{1}{3}h_{N_{\mathrm{e}}}.$$

Clearly, $\sum_{i\in\mathcal{A}_h} \overline{m}_i = |D|$ as expected.
(vi) It is possible to use the above basis together with the theory described in §81.1.2 to approximate hyperbolic systems because the coefficients of the lumped mass matrix are positive; see (81.5). This is the only required condition.
(viii) It is possible to apply Corollary 81.9 and Corollary 81.15 because the shape functions are nonnegative.

Exercise 82.4 (Gap estimates). (i) Let us denote by $\mathsf{U}^{\mathrm{L},n+1}$ the update given by (81.9) with the low-order graph viscosity d_{ij}^{L}. Subtracting (81.9) from (82.17), we obtain for all $i \in \mathcal{A}_h$,

$$\mathsf{U}_i^{n+1} = \mathsf{U}_i^{\mathrm{L},n+1} + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)}(d_{ij}^n - d_{ij}^{\mathrm{L},n})(\mathsf{U}_j^n - \mathsf{U}_i^n).$$

Introducing the auxiliary states $\overline{\mathsf{U}}_{ij}^n := \frac{1}{2}(\mathsf{U}_j^n + \mathsf{U}_i^n) - (\boldsymbol{f}(\mathsf{U}_j^n) - \boldsymbol{f}(\mathsf{U}_i^n))\cdot\frac{\boldsymbol{c}_{ij}}{2d_{ij}^{\mathrm{L},n}}$ as defined in the proof of Theorem 81.8, we have the identity (81.14), i.e.,

$$\mathsf{U}_i^{\mathrm{L},n+1} = \mathsf{U}_i^n\left(1 - \sum_{j\in\mathcal{I}(i)\setminus\{i\}}\frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\right) + \sum_{j\in\mathcal{I}(i)\setminus\{i\}}\frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\overline{\mathsf{U}}_{ij}^n.$$

An important property of the auxiliary states is that $\overline{\mathsf{U}}_{ij}^n \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$ (see Lemma 79.18 and Remark 79.19). Owing to the definition of γ_i^n and $d_{ii}^{\mathrm{L},n}$, we have $\gamma_i^n := \sum_{j\in\mathcal{I}(i)\setminus\{i\}}\frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}$, so that

$$\mathsf{U}_i^{*,n} := \frac{1}{\gamma_i^n}\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\overline{\mathsf{U}}_{ij}^n$$

is a convex combination of $\{\overline{\mathsf{U}}_{ij}^n\}_{i\in\mathcal{I}(i)}$. Hence, $\mathsf{U}_i^{*,n} \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$. Thus, we have $\mathsf{U}_i^{\mathrm{L},n+1} = (1-\gamma_i^n)\mathsf{U}_i^n + \gamma_i^n\mathsf{U}_i^{*,n}$, and this, in turn, implies that

$$\mathsf{U}_i^{n+1} = (1-\gamma_i^n)\mathsf{U}_i^n + \gamma_i^n\mathsf{U}_i^{*,n} + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}}(d_{ij}^n - d_{ij}^{\mathrm{L},n})(\mathsf{U}_j^n - \mathsf{U}_i^n).$$

(ii) Using that $\mathsf{U}_{ij}^{*,n} \leq \mathsf{U}_i^{\mathrm{M},n}$, we infer that

$$\mathsf{U}_i^{n+1} \leq \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n})(1-\gamma_i^n) + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}}(d_{ij}^n - d_{ij}^{\mathrm{L},n})(\mathsf{U}_j^n - \mathsf{U}_i^n).$$

Then, using that $d_{ij}^n \leq d_{ij}^{\mathrm{L},n}$ by definition, the above inequality gives

$$\begin{aligned}\mathsf{U}_i^{n+1} &\leq \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n})(1-\gamma_i^n) + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i^-)}(d_{ij}^{\mathrm{L},n} - d_{ij}^n)(\mathsf{U}_i^n - \mathsf{U}_j^n)\\ &\leq \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n})(1-\gamma_i^n) + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i^-)}(d_{ij}^{\mathrm{L},n} - d_{ij}^n)(\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n}).\end{aligned}$$

Now, using that $\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0$ and that U_i^n is in the convex hull of $\mathsf{U}_i^{\mathrm{M},n}$ and $\mathsf{U}_i^{\mathrm{m},n}$, we have $\mathsf{U}_i^n = \theta_i^n \mathsf{U}_i^{\mathrm{M},n} + (1-\theta_i^n)\mathsf{U}_i^{\mathrm{m},n}$, where $\theta_i^n \in [0,1]$ has been defined in (82.20). Hence, $\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n} = -\theta_i^n(\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})$ and $\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{M},n} = (1-\theta_i^n)(\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})$. With these definitions, the above inequality is rewritten

$$\mathsf{U}_i^{n+1} \leq \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})\Big((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n \frac{\tau}{m_i} \sum_{j \in \mathcal{I}(i^-)} (d_{ij}^{\mathrm{L},n} - d_{ij}^n)\Big).$$

(iii) Using that $d_{ij}^n \geq d_{ij}^{\mathrm{L},n}\psi_i^n$ and $\psi_i^n \geq 0$, we infer that $-d_{ij}^n \leq -d_{ij}^{\mathrm{L},n}\psi_i^n$, which, in turn, implies the following inequalities:

$$\begin{aligned}\mathsf{U}_i^{n+1} &\leq \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})\Big((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n(1-\psi_i^n)\frac{\tau}{m_i} \sum_{j \in \mathcal{I}(i^-)} d_{ij}^{\mathrm{L},n}\Big) \\ &= \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})\left((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n(1-\psi_i^n)\tfrac{1}{2}\gamma_i^{-,n}\right),\end{aligned}$$

by definition of $\gamma_i^{-,n}$.
(iv) The other estimate is obtained similarly. More precisely, using that $\mathsf{U}_i^{*,n} \geq \mathsf{U}_i^{\mathrm{m},n}$, we infer that

$$\begin{aligned}\mathsf{U}_i^{n+1} &\geq \mathsf{U}_i^{\mathrm{m},n} + (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n})(1-\gamma_i^n) + \frac{\tau}{m_i} \sum_{j \in \mathcal{I}(i^+)} (d_{ij}^n - d_{ij}^{\mathrm{L},n})(\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n) \\ &\geq \mathsf{U}_i^{\mathrm{m},n} + (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n})\left(\theta_i^n(1-\gamma_i^n) - (1-\psi_i^n)(1-\theta_i^n)\tfrac{1}{2}\gamma_i^{+,n}\right),\end{aligned}$$

by definition of $\gamma_i^{+,n}$.

Chapter 83

Higher-order approximation and limiting

Exercises

Exercise 83.1 (Dispersion error). Let $u(x,t)$ be a smooth function satisfying $\partial_t u + \beta \partial_x u = 0$, $x \in D := (0,1)$, $t > 0$, where $\beta \in \mathbb{R}$. Let $I \in \mathbb{N}\backslash\{0\}$ and consider the uniform mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$, $\forall i \in \{1{:}I{-}1\}$, with size $h := \frac{1}{I-1} = x_{i+1} - x_i$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions on $\mathcal{T}_h$ and let $\{\varphi_i\}_{i\in\mathcal{A}_h}$, $\mathcal{A}_h = \{1{:}I\}$, be the associated global Lagrange shape functions. (i) Compute the coefficients of the consistent mass matrix, $\mathcal{M}$, and the coefficients of the lumped mass matrix, $\overline{\mathcal{M}}$. (ii) Keep the time continuous and write the Galerkin approximation using the lumped mass matrix of the Cauchy problem (with the boundary condition equal to the initial condition as above) for a test function φ_i, $\forall i \in \mathcal{A}_h^\circ = \{2{:}I{-}1\}$. (iii) Let $\mathcal{I}_h^{\mathrm{L}}(u)$ be the Lagrange approximation of u. Using Taylor expansions, estimate (informally) the leading term in the consistency error $R_i^{\mathrm{L}}(t) := \frac{1}{\int_D \varphi_i \,\mathrm{d}x}\overline{\mathcal{M}}\partial_t u(x_i,t) + \int_D (\beta\partial_x \mathcal{I}_h^{\mathrm{L}}(u))\varphi_i \,\mathrm{d}x$, $\forall i \in \mathcal{A}_h^\circ$. (iv) Keep the time continuous and write the Galerkin approximation using the consistent mass matrix of the Cauchy problem for a test function φ_i, $\forall i \in \mathcal{A}_h^\circ$. (v) Using Taylor expansions, estimate (informally) the leading term in the consistency error $R_i(t) := \frac{1}{\int_D \varphi_i \,\mathrm{d}x}\int_D \big(\partial_t(\mathcal{I}_h^{\mathrm{L}}(u)) + \beta\partial_x(\mathcal{I}_h^{\mathrm{L}}(u))\big)\varphi_i \,\mathrm{d}x$, $\forall i \in \mathcal{A}_h^\circ$. (*Hint*: $u(x_i \pm h, t) = u(x_i) \pm h\partial_x u(x,t) + \frac{1}{2}h^2\partial_{xx}u(x_i,t) \pm \frac{1}{6}h^3\partial_{xxx}u(x_i,t) + \frac{1}{24}h^4\partial_{xxxx}u(x_i,t) \pm \frac{1}{120}h^5\partial_{xxxxx}u(x_i,t) + \mathcal{O}(h^6)$.)

Exercise 83.2 (FCT counterexample). Consider 1D Burgers' equation, $\boldsymbol{f}(u) := f(u)\boldsymbol{e}_x$, $f(u) := \frac{1}{2}u^2$, $D := (-1,1)$, with initial data $u_0(x) := -1$ if $x \le 0$ and $u_0(x) := 1$ otherwise. Let $I \ge 3$ be an odd number, and consider the (nonuniform) mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$, where the nodes x_i, $\forall i \in \mathcal{A}_h := \{1{:}I\}$, are such that $-1 =: x_1 < \cdots < x_I := 1$ and $x_{I'} \le 0 < x_{I'+1}$ with $I' := \frac{I+1}{2}$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions on $\mathcal{T}_h$ and let $\{\varphi_i\}_{i\in\mathcal{A}_h}$ be the associated global Lagrange shape functions. (i) Compute $\boldsymbol{c}_{i,i-1}$, $\boldsymbol{c}_{i,i}$, $\boldsymbol{c}_{i,i+1}$, and m_i for all $i \in \mathcal{A}_h^\circ := \{2{:}I{-}1\}$. (ii) Let $u_h^0 := \sum_{i\in\mathcal{A}_h} \mathsf{U}_i^0\varphi_i(x)$ with $\mathsf{U}_i^0 := -1$ if $i \le I'$ and $\mathsf{U}_i^0 := 1$ if $i > I'$. Compute the Galerkin solution at $t := \tau$ using the lumped mass matrix, say $u_h^{\mathrm{H},1}$. (iii) What is the maximum wave speed in the Riemann problem with the data $(-1,1)$? (iv) Compute the low-order solution at $t := \tau$, say $u_h^{\mathrm{L},1}$. (v) Using the notation of the FCT limiting, compute a_{ij} for all $i \in \mathcal{A}_h^\circ$ and all $j \in \mathcal{I}(i) := \{i-1, i, i+1\}$. (vi) Show that

$\ell_{ij} = 1$ for all $i \in \mathcal{A}_h^\circ$ and all $j \in \mathcal{I}(i)$. (vii) Does the approximate solution converge to the entropy solution?

Exercise 83.3 (Quasiconcavity). (i) Let $B \subset \mathbb{R}^m$ be a convex set. Show that a function $\Psi : B \to \mathbb{R}$ is quasiconcave iff for every finite set $\{\mathbf{U}_i\}_{i\in I} \subset B$ and all numbers $\{\theta_i\}_{i\in I} \subset [0,1]$ with $\sum_{i\in I} \theta_i = 1$, one has $\Psi(\sum_{i\in I} \theta_i \mathbf{U}_i) \geq \min_{i\in I} \Psi(\mathbf{U}_i)$. (ii) Let $\mathcal{A} \subset \mathbb{R}^m$ be a convex set. Let $\phi : \mathcal{A} \to \mathbb{R}$ be a quasiconcave function. Let $\boldsymbol{z} \in \mathbb{R}^m$, and let $L : \mathcal{A} \to \mathbb{R}$ be defined by $L(\boldsymbol{u}) := \boldsymbol{z}\cdot\boldsymbol{u}$ for all $\boldsymbol{u} \in \mathcal{A}$. Let $\phi : \mathcal{A} \to \mathbb{R}$ be a continuous function. Let $B := \{\boldsymbol{u} \in \mathcal{A} \mid L(\boldsymbol{u}) > 0\}$ and assume that $B \neq \emptyset$. Assume that $\psi : B \to \mathbb{R}$ defined by $\psi(\boldsymbol{u}) := L(\boldsymbol{u})\phi(\boldsymbol{u})$ is concave. Prove that $\phi_{|B} : B \to \mathbb{R}$ is quasiconcave. (A first example for the Euler equations is $B := \mathcal{A} = \{\boldsymbol{u} \in \mathbb{R}^m \mid \rho > 0\}$ with $L(\boldsymbol{u}) := \rho$, $\phi(\boldsymbol{u}) := e(\boldsymbol{u}) := \rho^{-1}E - \frac{1}{2}\rho^{-2}\boldsymbol{m}^2$, where $e(\boldsymbol{u})$ is the specific internal energy. Another example is $B := \mathcal{A} = \{\boldsymbol{u} \in \mathbb{R}^m \mid \rho > 0, e(\boldsymbol{u}) > 0\}$, $\phi(\boldsymbol{u}) := \Phi(\boldsymbol{u})$, where $\Phi(\boldsymbol{u})$ is the specific entropy.)

Exercise 83.4 (Harten's lemma). (i) Consider the following scheme for scalar conservation equations $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n - C_{i-1}^n(\mathsf{U}_i^n - \mathsf{U}_{i-1}^n) + D_i^n(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n)$ for all $i \in \mathbb{Z}$. Assume that $0 \leq C_i^n$, $0 \leq D_i^n$, $C_i^n + D_i^n \leq 1$ for all $i \in \mathbb{Z}$. Let $|\mathsf{V}|_{\mathrm{TV}} := \sum_{i\in\mathbb{Z}} |\mathsf{V}_{i+1} - \mathsf{V}_i|$ be the total variation of $\mathsf{V} \in \mathbb{R}^{\mathbb{Z}}$. Prove that the above algorithm is *total variation diminishing* (TVD), i.e., $|\mathsf{U}^{n+1}|_{\mathrm{TV}} \leq |\mathsf{U}^n|_{\mathrm{TV}}$. (ii) Consider the method described in (81.9)-(81.10) in dimension one. Assume that $\mathcal{I}(i) = \{i-1, i, i+1\}$ and that the mesh is infinite in both directions. Show that the method can be put into the above form and satisfies the above assumptions if $4\tau \sup_{i\in\mathbb{Z}} \frac{|d_{ii}^n|}{m_i} \leq 1$. (*Hint*: see Exercise 79.4.)

Exercise 83.5 (Lax–Wendroff). Let u be a smooth solution to the scalar transport equation $\partial_t u + a\partial_x u = 0$ with $a \in \mathbb{R}_+$. (i) Using finite Taylor expansions, show that $u(x, t_{n+1}) = u(x, t_n) - \tau a \partial_x u(x, t_n) + \frac{a^2\tau^2}{2}\partial_{xx} u(x, t_n) + \mathcal{O}(\tau^3)$. (ii) Consider now the time-stepping algorithm consisting of setting $u^0 := u_0$ and for all $n \geq 0$, $u^{n+1}(x) := u^n(x) - \tau a \partial_x u^n(x) + \frac{a^2\tau^2}{2}\partial_{xx} u^n(x)$. What is the (informal) order of accuracy of this method with respect to τ? (iii) Let $\mathcal{T}_h$ be a uniform mesh in $D := (0,1)$ with grid points $x_i := (i-1)h$, $\forall i \in \mathcal{A}_h := \{1{:}I\}$, $h := \frac{1}{I-1}$. Let $\{\varphi_i\}_{i\in\mathcal{A}_h}$ be the piecewise linear Lagrange shape functions associated with the grid points $\{x_i\}_{i\in\mathcal{A}_h}$. Let x_i be an interior node, i.e., $i \in \mathcal{A}_h^\circ := \{2{:}I{-}1\}$. Write the equation corresponding to the Galerkin approximation using the lumped mass matrix of the equation $u^{n+1}(x) = u^n(x) - \tau a \partial_x u^n(x) + \frac{a^2\tau^2}{2}\partial_{xx} u^n(x)$ with homogeneous Neumann boundary conditions using the test function φ_i, where both u^{n+1} and u^n are approximated in $P_1^{\mathrm{g}}(\mathcal{T}_h) := \mathrm{span}\{\varphi_i\}_{i\in\mathcal{A}_h}$. (iv) What is the (informal) order of accuracy of this method with respect to τ and h? (v) Let $u_h^{\mathrm{L},n+1} := \sum_{i\in\mathcal{A}_h} \mathsf{U}_i^{\mathrm{L},n+1}\varphi_i$ be the first-order approximation of u using (81.9)-(81.10). Show that $m_i\mathsf{U}_i^{n+1} = m_i\mathsf{U}_i^{\mathrm{L},n+1} + \frac{a\tau}{2}(\lambda - 1)(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n) + \frac{a\tau}{2}(\lambda - 1)(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n)$, where $\gamma := \frac{a\tau}{h}$. *Note*: the scheme is now ready for FCT limiting. Actually, there exists in the literature a plethora of limiting techniques (like FCT) that, after applying the limiter, make the scheme TVD in the sense of Exercise 83.4; see Sweby [42].

Solution to exercises

Exercise 83.1 (Dispersion error). (i) Let $m_{ij} := \int_D \varphi_i \varphi_j \,\mathrm{d}x$ be the coefficients of the consistent mass matrix $\mathcal{M}$ for all $i, j \in \mathcal{A}_h := \{1{:}I\}$. We have

$$\begin{aligned} m_{ij} &= 0 \quad \text{if } |i-j| \geq 2, \\ m_{11} &= \frac{1}{3}h, \quad m_{12} = \frac{1}{6}h, \\ m_{ii-1} &= \frac{1}{6}h, \quad m_{ii} = \frac{4}{6}h, \quad m_{ii+1} = \frac{1}{6}h, \quad \forall i \in \mathcal{A}_h^\circ = \{2{:}I{-}1\}, \\ m_{I,I+1} &= \frac{1}{6}h, \quad m_{I+1,I+1} = \frac{1}{3}h. \end{aligned}$$

Let $m_i := \int_D \varphi_i \,\mathrm{d}x = \sum_{j\in\mathcal{A}_h} m_{ij}$ be the diagonal coefficients of the lumped mass matrix $\overline{\mathcal{M}}$ for all $i \in \mathcal{A}_h$ (all the off-diagonal coefficients are zero). We have

$$m_1 = \frac{1}{2}h, \quad m_i = h, \ \forall i \in \mathcal{A}_h^\circ, \quad m_{I+1} = \frac{1}{2}h.$$

(ii) We have $\mathcal{I}(i) = \{i-1, i, i+1\}$ for all $i \in \mathcal{A}_h^\circ$, and recalling Example 81.5, we have $\boldsymbol{c}_{i,i-1} = -\frac{1}{2}\boldsymbol{e}_x$ and $\boldsymbol{c}_{i,i+1} = \frac{1}{2}\boldsymbol{e}_x$, where $\boldsymbol{e}_x$ is the unit vector orienting $\mathbb{R}$. The Galerkin approximation of the Cauchy problem using the lumped mass matrix $\overline{\mathcal{M}}$ is formulated as follows: Find $u_h(t) := \sum_{i\in\mathcal{A}_h} \mathsf{U}_j(t)\varphi_j$ such that

$$h\partial_t \mathsf{U}_i(t) + \frac{1}{2}(\mathsf{U}_{i+1}(t) - \mathsf{U}_{i-1}(t)) = 0, \qquad \forall i \in \mathcal{A}_h^\circ,$$

with $\mathsf{U}_1(t)$ and $\mathsf{U}_I(t)$ prescribed by the boundary condition coming from the initial condition.
(iii) By definition, we have

$$R_i^{\mathrm{L}}(t) = h\partial_t u(x_i,t) + \frac{1}{2}(u(x_{i+1},t) - u(x_{i-1},t)),$$

for all $i \in \mathcal{A}_h^\circ$. Using $x_{i\pm1} = x_i \pm h$ and the Taylor expansion

$$u(x_i \pm h, t) = u(x_i) \pm h\partial_x u(x,t) + \frac{1}{2}h^2\partial_{xx}u(x_i,t) \pm \frac{1}{6}h^3\partial_{xxx}u(x_i,t) + \frac{1}{24}h^4\partial_{xxxx}u(x_i,t) + \mathcal{O}(h^5),$$

we infer that

$$R_i^{\mathrm{L}}(t) = (\partial_t u + \beta\partial_x u)(x_i,t) + \beta\frac{h^2}{6}\partial_{xxx}u(x_i,t) + \mathcal{O}(h^4).$$

In conclusion, we have $R_i^{\mathrm{L}}(t) = \beta\frac{h^2}{6}\partial_{xxx}u(x_i,t) + \mathcal{O}(h^4)$. The leading term of the consistency error at x_i is second-order in h and proportional to a third-order partial derivative of u with respect to x.
(iv) The Galerkin approximation of the Cauchy problem using the consistent mass matrix $\mathcal{M}$ is formulated as follows: Find $v_h(t) := \sum_{i\in\mathcal{A}_h} \mathsf{U}_j(t)\varphi_j$ such that

$$\frac{1}{6}h\big(\partial_t \mathsf{U}_{i-1}(t) + 4\partial_t \mathsf{U}_i(t) + \partial_t \mathsf{U}_{i+1}(t)\big) + \frac{1}{2}(\mathsf{U}_{i+1}(t) - \mathsf{U}_{i-1}(t)) = 0,$$

for all $i \in \mathcal{A}_h^\circ$.
(v) Using the definition of the mass matrix, we have

$$\frac{1}{h}\sum_{j\in\{i-1:\,i+1\}} m_{ij}\partial_t u(x_j,t) = \partial_t u(x_i,t) + \frac{1}{6}\big(\partial_t u(x_{i-1},t) - 2\partial_t u(x_i,t) + \partial_t u(x_{i+1},t)\big).$$

Using Taylor expansions shows that

$$\begin{aligned}\partial_t u(x_i \pm h,t) = \partial_t u(x_i) \pm h\partial_{xt}u(x,t) + \frac{1}{2}h^2\partial_{xxt}u(x_i,t)\\ \pm \frac{1}{6}h^3\partial_{xxxt}u(x_i,t) + \frac{1}{24}h^4\partial_{xxxxt}u(x_i,t) + \mathcal{O}(h^4),\end{aligned}$$

whence we infer that

$$\begin{aligned}\frac{1}{h}\sum_{j\in\{i-1:\,i+1\}} m_{ij}\partial_t u(x_j,t) &= \partial_t u(x_i,t) + \frac{h^2}{6}\partial_{txx}u(x_i,t) + \frac{h^4}{72}\partial_{txxxx}u(x_i,t) + \mathcal{O}(h^6)\\ &= \partial_t u(x_i,t) - \beta\frac{h^2}{6}\partial_{xxx}u(x_i,t) - \beta\frac{h^4}{72}\partial_{xxxxx}u(x_i,t) + \mathcal{O}(h^6).\end{aligned}$$

By using again that

$$\begin{aligned}u(x_i \pm h,t) = u(x_i) \pm h\partial_x u(x,t) + \frac{1}{2}h^2\partial_{xx}u(x_i,t) \pm \frac{1}{6}h^3\partial_{xxx}u(x_i,t)\\ + \frac{1}{24}h^4\partial_{xxxx}u(x_i,t) \pm \frac{1}{120}h^5\partial_{xxxxx}u(x_i,t) + \mathcal{O}(h^6),\end{aligned}$$

we infer that

$$\begin{aligned}\frac{1}{h}\sum_{j\in\{i-1:\,i+1\}} m_{ij}\partial_t u(x_j,t) + \beta\frac{u(x_{i+1},t) - u(x_{i-1},t)}{2h}\\ = \partial_t u(x_i,t) + \beta\partial_x u(x_i,t) - \beta\frac{1}{180}h^4\partial_{xxxxx}u(x_i,t) + \mathcal{O}(h^6).\end{aligned}$$

This shows that $R_i(u) = \beta\frac{1}{180}h^4\partial_{xxxxx}u(x_i,t) + \mathcal{O}(h^6)$. The consistency error is fourth-order in h at the interior grid points. This means that the Galerkin approximation using the consistent mass matrix is superconvergent at the interior grid points, which is not the case when the lumped mass matrix is used.

Exercise 83.2 (FCT counterexample). (i) We have $\boldsymbol{c}_{i,i-1} = -\frac{1}{2}\boldsymbol{e}_x$, $\boldsymbol{c}_{i,i} = \boldsymbol{0}$ and $\boldsymbol{c}_{i,i+1} = \frac{1}{2}\boldsymbol{e}_x$, $m_i = \frac{h_i+h_{i+1}}{2}$. The equation for U_i^{n+1} is

$$\mathsf{U}_i^{n+1} = \mathsf{U}_i^n + \frac{\tau}{2m_i}(f(\mathsf{U}_{i-1}^n) - f(\mathsf{U}_{i+1}^n)) + \frac{\tau}{m_i}d_{i,i-1}^n(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n) + \frac{\tau}{m_i}d_{i,i+1}^n(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n),$$

with the convention that $\mathsf{U}_1^n := -1$ and $\mathsf{U}_I^n := 1$.
(ii) Let $\mathsf{U}^{\mathrm{H},1}$ be the Galerkin solution at $t_1 := \tau$ which is, by definition, obtained by solving the above equation with $d_{ij}^1 = 0$. Since $f(\mathsf{U}_{i-1}^0) - f(\mathsf{U}_{i+1}^0) = 0$ for all $i \in \{0{:}2N\}$, we obtain

$$\mathsf{U}^{\mathrm{H},1} = \mathsf{U}^0.$$

(iii) The entropy solution is an expansion wave. The maximum wave speed is $\max(f'(-1), f'(1)) = 1$. See also Example 79.17.

(iv) Let us compute the low-order solution $\mathsf{U}^{\mathrm{L},1}$. It is clear that $\mathsf{U}_i^{\mathrm{L},1} = \mathsf{U}_i^0$ for all $i < I'$ and all $i > I'+1$. We have $\mathsf{U}_{I'-1}^0 = -1$, $\mathsf{U}_{I'}^0 = -1$, $\mathsf{U}_{I'+1}^0 = 1$, and $f(\mathsf{U}_{i+1}^0) - f(\mathsf{U}_{i-1}^0) = 0$ for all $i \in \{I', I'+1\}$. Note also that $d_{I',I'+1}^{\mathrm{L},0} = d_{I'+1,I'}^{\mathrm{L},0} = \frac{1}{2}$ since the maximum wave speed in the Riemann problem with the data $(-1,1)$ is 1. We infer that

$$\mathsf{U}_{I'}^{\mathrm{L},1} = \mathsf{U}_{I'}^0 + \frac{\tau}{m_I'} d_{I',I'+1}(\mathsf{U}_{I'+1}^0 - \mathsf{U}_{I'}^0) = -1 + \frac{\tau}{m_I'},$$
$$\mathsf{U}_{I'+1}^{\mathrm{L},1} = \mathsf{U}_{I'+1}^0 + \frac{\tau}{m_{I'+1}} d_{I'+1,I'}(\mathsf{U}_{I'}^0 - \mathsf{U}_{I'+1}^0) = 1 - \frac{\tau}{m_{I'+1}}.$$

(v) In the FCT notation, we have

$$m_{I'}\mathsf{U}_{I'}^{\mathrm{H},1} = m_{I'}\mathsf{U}_{I'}^{\mathrm{L},1} - \frac{\tau}{2}(\mathsf{U}_{I'+1}^0 - \mathsf{U}_{I'}^0)$$
$$m_{I'+1}\mathsf{U}_{I'+1}^{\mathrm{H},1} = m_{I'+1}\mathsf{U}_{I'+1}^{\mathrm{L},1} - \frac{\tau}{2}(\mathsf{U}_{I'}^0 - \mathsf{U}_{I'+1}^0).$$

This means that $a_{I',I'+1} = -\frac{\tau}{2}(\mathsf{U}_{I'+1}^0 - \mathsf{U}_{I'}^0) = -\tau$ and $a_{I'+1,I'} = -\frac{\tau}{2}(\mathsf{U}_{I'}^0 - \mathsf{U}_{I'+1}^0) = \tau$.
(vi) Let us now compute the limiter coefficient $\ell_{I',I'+1}$ with $\mathsf{U}_{I'}^{\max} = \mathsf{U}_{I'+1}^{\max} = 1$ and $\mathsf{U}_{I'}^{\min} = \mathsf{U}_{I'+1}^{\min} = -1$. We evaluate the FCT coefficients as follows:

$$P_{I'}^+ = 0, \qquad P_{I'}^- = -\tau, \qquad P_{I'+1}^+ = \tau, \qquad P_{I'+1}^- = 0,$$
$$Q_{I'}^+ = 2m_{I'} - \tau, \qquad Q_{I'}^- = -\tau, \qquad Q_{I'+1}^+ = \tau, \qquad Q_{I'+1}^- = -2m_{I'+1} + \tau,$$
$$R_{I'}^+ = 1, \qquad R_{I'}^- = 1, \qquad R_{I'+1}^+ = 1, \qquad R_{I'+1}^- = 1,$$

which gives $\ell_{I',I'+1} = 1$. Hence, $\mathsf{U}^1 = \mathsf{U}^{\mathrm{H},1}$, so that $\mathsf{U}^1 = \mathsf{U}^0$ since $\mathsf{U}^{\mathrm{H},1} = \mathsf{U}^0$. (vii) In conclusion, $u_h^1 = u_h^0$, i.e., $u_h^n = u_h^0$ for all $n \geq 0$. This proves that the numerical solution is a stationary discontinuity, whereas the entropy solution of the problem is an expansion wave. Hence, the method does not converge to the entropy solution.

Exercise 83.3 (Quasiconcavity). (i) Assume that $\Psi : B \to \mathbb{R}$ is quasiconcave. Let $\{\mathbf{U}_i\}_{i\in I} \subset B$ and $\{\theta_i\}_{i\in I} \subset [0,1]$ with $\sum_{i\in I}\theta_i = 1$. Taking $\lambda := \min_{i\in I}\Psi(\mathbf{U}_i)$, the upper level set $L_\lambda(\Psi)$ is convex. Since $\mathbf{U}_i \in L_\lambda(\Psi)$ for all $i \in I$, we infer that $\sum_{i\in I}\theta_i\mathbf{U}_i \in L_\lambda(\Psi)$. This proves that $\Psi(\sum_{i\in I}\theta_i\mathbf{U}_i) \geq \lambda = \min_{i\in I}\Psi(\mathbf{U}_i)$. Conversely, assume that for all $\{\mathbf{U}_i\}_{i\in I} \subset B$ and $\{\theta_i\}_{i\in I} \subset [0,1]$ with $\sum_{i\in I}\theta_i = 1$, one has $\Psi(\sum_{i\in I}\theta_i\mathbf{U}_i) \geq \min_{i\in I}\Psi(\mathbf{U}_i)$. Let $\lambda \in \mathbb{R}$ and consider the upper level set $L_\lambda(\Psi)$. If $L_\lambda(\Psi)$ is empty, there is nothing to prove. Otherwise, let $\mathbf{U}_1, \mathbf{U}_2 \in L_\lambda(\Psi)$ and let $t \in [0,1]$. Then our assumption with $I := \{1,2\}$, $\theta_1 := t$, $\theta_2 := 1-t$ implies that $\Psi(t\mathbf{U}_1 + (1-t)\mathbf{U}_2) \geq \min(\Psi(\mathbf{U}_1), \Psi(\mathbf{U}_2)) \geq \lambda$. Hence, $t\mathbf{U}_1 + (1-t)\mathbf{U}_2 \in L_\lambda(\Psi)$. This proves the convexity of $L_\lambda(\Psi)$, and therefore the quasiconcavity of Ψ.
(ii) Let $\lambda \in \mathbb{R}$ and $L_\lambda := \{\boldsymbol{u} \in B \mid \phi(\boldsymbol{u}) \geq \lambda\}$ and $G_\lambda := \{\boldsymbol{u} \in B \mid \psi(\boldsymbol{u}) - \lambda L(\boldsymbol{u}) \geq 0\}$. Let $\boldsymbol{u} \in L_\lambda$. We have $\psi(\boldsymbol{u}) = L(\boldsymbol{u})\phi(\boldsymbol{u}) \geq L(\boldsymbol{u})\lambda$ because $L(\boldsymbol{u}) > 0$. Hence, $\boldsymbol{u} \in G_\lambda$. Conversely, let $\boldsymbol{u} \in G_\lambda$. Then $\psi(\boldsymbol{u}) = L(\boldsymbol{u})\phi(\boldsymbol{u}) \geq L(\boldsymbol{u})\lambda$ implies that $\phi(\boldsymbol{u}) \geq \lambda$ because $L(\boldsymbol{u}) > 0$ (recall that $\boldsymbol{u} \in G_\lambda$ implies that $\boldsymbol{u} \in B$). This proves that $L_\lambda = G_\lambda$. The function $\psi(\boldsymbol{u}) - \lambda L(\boldsymbol{u})$ is concave since $\psi(\boldsymbol{u})$ is concave and L is linear. Hence, G_λ is convex since it is the zero upper level set of $\psi(\boldsymbol{u}) - \lambda L(\boldsymbol{u})$ and B is convex. This proves that L_λ is convex for all $\lambda \in \mathbb{R}$. Hence, $\phi_{|B}$ is quasiconcave.

Exercise 83.4 (Harten's lemma). We have

$$\mathsf{U}_i^{n+1} = \mathsf{U}_i^n - C_{i-1}^n(\mathsf{U}_i^n - \mathsf{U}_{i-1}^n) + D_i^n(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n),$$
$$\mathsf{U}_{i+1}^{n+1} = \mathsf{U}_{i+1}^n - C_i^n(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n) + D_{i+1}^n(\mathsf{U}_{i+2}^n - \mathsf{U}_{i+1}^n).$$

Taking the difference, we obtain

$$\begin{aligned}\mathsf{U}_{i+1}^{n+1}-\mathsf{U}_i^{n+1}=\mathsf{U}_{i+1}^n-\mathsf{U}_i^n-C_i^n(\mathsf{U}_{i+1}^n-\mathsf{U}_i^n)+C_{i-1}^n(\mathsf{U}_i^n-\mathsf{U}_{i-1}^n)\\+D_{i+1}^n(\mathsf{U}_{i+2}^n-\mathsf{U}_{i+1}^n)-D_i^n(\mathsf{U}_{i+1}^n-\mathsf{U}_i^n).\end{aligned}$$

Rearranging the terms, we infer that

$$\mathsf{U}_{i+1}^{n+1}-\mathsf{U}_i^{n+1}=(\mathsf{U}_{i+1}^n-\mathsf{U}_i^n)(1-C_i^n-D_i^n)+C_{i-1}^n(\mathsf{U}_i^n-\mathsf{U}_{i-1}^n)+D_{i+1}^n(\mathsf{U}_{i+2}^n-\mathsf{U}_{i+1}^n).$$

We take the absolute value on both sides and use the given assumptions $0\le C_i^n$, $0\le D_i^n$, $C_i^n+D_i^n\le 1$ to infer that

$$|\mathsf{U}_{i+1}^{n+1}-\mathsf{U}_i^{n+1}|\le|\mathsf{U}_{i+1}^n-\mathsf{U}_i^n|(1-C_i^n-D_i^n)+C_{i-1}^n|\mathsf{U}_i^n-\mathsf{U}_{i-1}^n|+D_{i+1}^n|\mathsf{U}_{i+2}^n-\mathsf{U}_{i+1}^n|.$$

Summing over the index $i\in\mathbb{Z}$, we obtain

$$\begin{aligned}\sum_{i\in\mathbb{Z}}|\mathsf{U}_{i+1}^{n+1}-\mathsf{U}_i^{n+1}|&\le\sum_{i\in\mathbb{Z}}|\mathsf{U}_{i+1}^n-\mathsf{U}_i^n|(1-C_i^n-D_i^n)+\sum_{i\in\mathbb{Z}}C_{i-1}^n|\mathsf{U}_i^n-\mathsf{U}_{i-1}^n|\\&\quad+\sum_{i\in\mathbb{Z}}D_{i+1}^n|\mathsf{U}_{i+2}^n-\mathsf{U}_{i+1}^n|\\&\le\sum_{i\in\mathbb{Z}}|\mathsf{U}_{i+1}^n-\mathsf{U}_i^n|(1-C_i^n-D_i^n)+\sum_{i\in\mathbb{Z}}C_i^n|\mathsf{U}_{i+1}^n-\mathsf{U}_i^n|\\&\quad+\sum_{i\in\mathbb{Z}}D_i^n|\mathsf{U}_{i+1}^n-\mathsf{U}_i^n|\\&=\sum_{i\in\mathbb{Z}}|\mathsf{U}_{i+1}^n-\mathsf{U}_i^n|.\end{aligned}$$

This proves that $\sum_{i\in\mathbb{Z}}|\mathsf{U}_{i+1}^{n+1}-\mathsf{U}_i^{n+1}|\le\sum_{i\in\mathbb{Z}}|\mathsf{U}_{i+1}^n-\mathsf{U}_i^n|$.
(ii) Recalling that $\boldsymbol{c}_{i,i-1}+\boldsymbol{c}_{i,i}+\boldsymbol{c}_{i,i+1}=\mathbf{0}$ and that $d_{i,i-1}+d_{i,i}+d_{i,i+1}=0$, the scheme (81.9)-(81.10) can be put into the following form:

$$\begin{aligned}\mathsf{U}_i^{n+1}&=\mathsf{U}_i^n+\tfrac{\tau}{m_i}\big((\boldsymbol{f}(\mathsf{U}_i^n)-\boldsymbol{f}(\mathsf{U}_{i-1}^n))\cdot\boldsymbol{c}_{i,i-1}-d_{i,i-1}^n(\mathsf{U}_i^n-\mathsf{U}_{i-1}^n)\big)\\&\quad+\tfrac{\tau}{m_i}\big((\boldsymbol{f}(\mathsf{U}_i^n)-\boldsymbol{f}(\mathsf{U}_{i+1}^n))\cdot\boldsymbol{c}_{i,i+1}+d_{i,i+1}^n(\mathsf{U}_{i+1}^n-\mathsf{U}_i^n)\big)\\&=\mathsf{U}_i^n-\tfrac{\tau}{m_i}\big(-\tfrac{\boldsymbol{f}(\mathsf{U}_i^n)-\boldsymbol{f}(\mathsf{U}_{i-1}^n)}{\mathsf{U}_i^n-\mathsf{U}_{i-1}^n}\cdot\boldsymbol{c}_{i,i-1}+d_{i,i-1}^n\big)(\mathsf{U}_i^n-\mathsf{U}_{i-1}^n)\\&\quad+\tfrac{\tau}{m_i}\big(\tfrac{\boldsymbol{f}(\mathsf{U}_i^n)-\boldsymbol{f}(\mathsf{U}_{i+1}^n)}{\mathsf{U}_{i+1}^n-\mathsf{U}_i^n}\cdot\boldsymbol{c}_{i,i+1}+d_{i,i+1}^n\big)(\mathsf{U}_{i+1}^n-\mathsf{U}_i^n)\big).\end{aligned}$$

Thus, $\mathsf{U}_i^{n+1}=\mathsf{U}_i^n-C_{i-1}^n(\mathsf{U}_i^n-\mathsf{U}_{i-1}^n)+D_i^n(\mathsf{U}_{i+1}^n-\mathsf{U}_i^n)$ with

$$\begin{aligned}C_{i-1}^n&:=\tfrac{\tau}{m_i}\big(-\tfrac{\boldsymbol{f}(\mathsf{U}_i^n)-\boldsymbol{f}(\mathsf{U}_{i-1}^n)}{\mathsf{U}_i^n-\mathsf{U}_{i-1}^n}\cdot\boldsymbol{c}_{i,i-1}+d_{i,i-1}^n\big),\\D_i^n&:=\tfrac{\tau}{m_i}\big(-\tfrac{\boldsymbol{f}(\mathsf{U}_{i+1}^n)-\boldsymbol{f}(\mathsf{U}_i^n)}{\mathsf{U}_{i+1}^n-\mathsf{U}_i^n}\cdot\boldsymbol{c}_{i,i+1}+d_{i,i+1}^n\big).\end{aligned}$$

Recall that $\boldsymbol{c}_{ij}:=\boldsymbol{e}_x\int_D\varphi_i\partial_x\varphi_j\,\mathrm{d}x$. Let us set $\boldsymbol{f}(v):=f(v)\boldsymbol{e}_x$ and $\boldsymbol{n}_{ij}:=\boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$. Recalling Exercise 79.4, we know that $\lambda_{\max}(\boldsymbol{n}_{ij},\mathsf{U}_i^n,\mathsf{U}_{i-1}^n)\ge|\frac{f(\mathsf{U}_i^n)-f(\mathsf{U}_{i-1}^n)}{\mathsf{U}_i^n-\mathsf{U}_{i-1}^n}|$. Thus, we have

$$d_{i,i-1}^n\ge\|\boldsymbol{c}_{i,i-1}\|\lambda_{\max}(\boldsymbol{n}_{ij},\mathsf{U}_i^n,\mathsf{U}_{i-1}^n)\ge\|\boldsymbol{c}_{i,i-1}\|\times\Big|\tfrac{f(\mathsf{U}_i^n)-f(\mathsf{U}_{i-1}^n)}{\mathsf{U}_i^n-\mathsf{U}_{i-1}^n}\Big|.$$

Hence, $C_{i-1}^n \geq 0$. We prove similarly that $D_i^n \geq 0$. The same argument shows that

$$C_i^n + D_i^n \leq 2\tfrac{\tau}{m_i}(d_{i+1,i}^n + d_{i,i+1}^n) = 4\tfrac{\tau}{m_i} d_{i+1,i}^n \leq 4\tfrac{\tau}{m_i}|d_{i,i}^n| \leq 1.$$

Exercise 83.5 (Lax–Wendroff). (i) Let us start by observing that $\partial_t u = -a\partial u$ and $\partial_{tt} u = -a\partial_x(\partial_t u) = a^2\partial_{xx} u$. Using a finite Taylor expansion with respect to t, we infer that

$$\begin{aligned} u(x,t_{n+1}) &= u(x,t_n) + \tau\partial_t u(x,t_n) + \tfrac{\tau^2}{2}\partial_{tt} u(x,t_n) + \mathcal{O}(\tau^3) \\ &= u(x,t_n) - \tau a\partial_x u(x,t_n) + \tfrac{a^2\tau^2}{2}\partial_{xx} u(x,t_n) + \mathcal{O}(\tau^3). \end{aligned}$$

(ii) The local truncation error is $\mathcal{O}(\tau^3)$, but after $\frac{1}{\tau}$ time steps, the error is (informally) $\mathcal{O}(\tau^2)$. Hence, the scheme is (informally) second-order accurate in time.
(iii) Let us set $u_h^n(x) := \sum_{i\in\mathcal{A}_h} \mathsf{U}_i^n \varphi_i$ and $u_h^{n+1}(x) := \sum_{i\in\mathcal{A}_h} \mathsf{U}_i^{n+1}\varphi_i$. We observe that $m_i := \int_D \varphi_i \,\mathrm{d}x = h$. After integrating by parts the second-order derivative, we obtain

$$\begin{aligned} m_i\mathsf{U}_i^{n+1} &= m_i\mathsf{U}_i^n - a\tau \sum_{i\in\mathcal{A}_h} \mathsf{U}_j^n \int_D \varphi_i\partial_x\varphi_j \,\mathrm{d}x - \frac{a^2\tau^2}{2} \sum_{i\in\mathcal{A}_h} \mathsf{U}_j^n \int_D \partial_x\varphi_i\partial_x\varphi_j \,\mathrm{d}x \\ &= m_i\mathsf{U}_i^n - \frac{a\tau}{2}(\mathsf{U}_{i+1}^n - \mathsf{U}_{i-1}^n) + \frac{a^2\tau^2}{2h}(\mathsf{U}_{i+1}^n - 2\mathsf{U}_i^n + \mathsf{U}_{i-1}^n). \end{aligned}$$

The boundary terms have been removed to account for the homogeneous Neumann boundary conditions.
(iv) Since we did not change anything on the time stepping, the (informal) accuracy in time is τ^2. The approximation in space being exact for linear solutions, the (informal) order of accuracy in space is h^2. Hence, the method is (informally) second-order accurate.
(v) Using the computation done in Example 81.5, we obtain

$$m_i\mathsf{U}_i^{\mathrm{L},n+1} = m_i\mathsf{U}_i^n - a\tau(\mathsf{U}_i^n - \mathsf{U}_{i-1}^n).$$

Upon introducing the quantity $\gamma = \frac{a\tau}{h}$, we infer that

$$\begin{aligned} m_i\mathsf{U}_i^{n+1} &= m_i\mathsf{U}_i^n - a\tau(\mathsf{U}_i^n - \mathsf{U}_{i-1}^n) - \frac{a\tau}{2}(\mathsf{U}_{i+1}^n - 2\mathsf{U}_i^n + \mathsf{U}_{i-1}^n) + \frac{a^2\tau^2}{2h}(\mathsf{U}_{i+1}^n - 2\mathsf{U}_i^n + \mathsf{U}_{i-1}^n) \\ &= m_i\mathsf{U}_i^{\mathrm{L},n+1} + \frac{a\tau}{2}(\lambda-1)(\mathsf{U}_{i+1}^n - 2\mathsf{U}_i^n + \mathsf{U}_{i-1}^n) \\ &= m_i\mathsf{U}_i^{\mathrm{L},n+1} + \frac{a\tau}{2}(\lambda-1)(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n) + \frac{a\tau}{2}(\lambda-1)(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n). \end{aligned}$$

Bibliography

[1] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus, homological techniques, and applications. *Acta Numer.*, 15:1–155, 2006. pages 83

[2] I. Babuška. The finite element method with Lagrangian multipliers. *Numer. Math.*, 20: 179–192, 1973. pages 271

[3] H. Beirão da Veiga. On a stationary transport equation. *Ann. Univ. Ferrara Sez. VII*, 32: 79–91, 1986. pages 164, 321

[4] S. Bertoluzza. The discrete commutator property of approximation spaces. *C. R. Acad. Sci. Paris, Sér. I*, 329(12):1097–1102, 1999. pages 110

[5] J. H. Bramble and S. R. Hilbert. Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation. *SIAM J. Numer. Anal.*, 7:112–124, 1970. pages 111

[6] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, NY, 2011. pages 341

[7] F. Brezzi and K.-J. Bathe. A discourse on the stability conditions for mixed finite element formulations. *Comput. Methods Appl. Mech. Engrg.*, 82(1-3):27–57, 1990. pages 292

[8] E. Burman. Robust error estimates in weak norms for advection dominated transport problems with rough data. *Math. Models Methods Appl. Sci.*, 24(13):2663–2684, 2014. pages 164, 321

[9] J. C. Butcher. Implicit Runge-Kutta processes. *Math. Comp.*, 18:50–64, 1964. pages 372

[10] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. A posteriori error control for DPG methods. *SIAM J. Numer. Anal.*, 52(3):1335–1353, 2014. pages 264

[11] L. Chesnel and P. Ciarlet, Jr. T-coercivity and continuous Galerkin methods: application to transmission problems with sign changing coefficients. *Numer. Math.*, 124(1):1–29, 2013. pages 126

[12] J.-P. Croisille. Finite volume box schemes and mixed methods. *M2AN Math. Model. Numer. Anal.*, 34(2):1087–1106, 2000. pages 272

[13] F. Demengel and G. Demengel. *Functional spaces for the theory of elliptic partial differential equations*. Universitext. Springer, London, UK; EDP Sciences, Les Ulis, France, 2012. Translated from the 2007 French original by Reinie Erné. pages 16

[14] L. F. Demkowicz and J. Gopalakrishnan. An overview of the discontinuous Petrov Galerkin method. In *Recent developments in discontinuous Galerkin finite element methods for partial differential equations*, volume 157 of *The IMA Volumes in Mathematics and its Applications*, pages 149–180. Springer, Cham, Switzerland, 2014. pages 264

[15] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications [Mathematics & Applications]*. Springer-Verlag, Berlin, 2012. pages 197

[16] N. Dyn, D. Levine, and J. A. Gregory. A butterfly subdivision scheme for surface interpolation with tension control. *ACM Trans. Graph.*, 9(2):160–169, Apr. 1990. pages 68

[17] M. Fortin and M. Soulié. A non-conforming piecewise quadratic finite element on triangles. *Internat. J. Numer. Methods Engrg.*, 19:505–520, 1983. pages 190

[18] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Math. Comp.*, 83 (286):537–552, 2014. pages 264

[19] I. Greff. *Schémas boîte : étude théorique et numérique*. PhD thesis, University of Metz, France, 2003. pages 190

[20] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985. pages 11, 193

[21] J.-L. Guermond. Remarques sur les méthodes de projection pour l'approximation des équations de Navier-Stokes. *Numer. Math.*, 67(4):465–473, 1994. pages 393

[22] J.-L. Guermond and A. Salgado. A note on the Stokes operator and its powers. *J. Appl. Math. Comput.*, 36(1-2):241–250, 2011. pages 393

[23] K. Gustafson. The Toeplitz-Hausdorff theorem for linear operators. *Proc. Amer. Math. Soc.*, 25:203–204, 1970. pages 238

[24] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations. II. Stiff and Differential-algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2010. Second revised edition, paperback. pages 372

[25] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I. Nonstiff problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993. pages 372

[26] A. Harten, P. D. Lax, C. D. Levermore, and W. J. Morokoff. Convex entropies and hyperbolicity for general Euler equations. *SIAM J. Numer. Anal.*, 35(6):2117–2127, 1998. pages 433

[27] B. Holm and T. P. Wihler. Continuous and discontinuous Galerkin time stepping methods for nonlinear initial value problems with application to finite time blow-up. *Numer. Math.*, 138(3):767–799, 2018. pages 363, 366

[28] C. O. Horgan. Korn's inequalities and their applications in continuum mechanics. *SIAM Rev.*, 37:491–511, 1995. pages 222

[29] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number. I. The h-version of the FEM. *Comput. Math. Appl.*, 30(9):9–37, 1995. pages 185

[30] L. John, M. Neilan, and I. Smears. Stable discontinuous Galerkin FEM without penalty parameters. In *Numerical Mathematics and Advanced Applications ENUMATH 2015*, volume 112 of *Lecture Notes in Computational Science and Engineering*, pages 165–173. Springer, Cham, Switzerland, 2016. pages 204

[31] C. Johnson and A. Szepessy. On the convergence of a finite element method for a nonlinear hyperbolic conservation law. *Math. Comp.*, 49(180):427–444, 1987. pages 110

[32] K. Y. Kim. Guaranteed a posteriori error estimator for mixed finite element methods of linear elasticity with weak stress symmetry. *SIAM J. Numer. Anal.*, 48(6):2364–2385, 2011. pages 222

[33] L. D. Marini. An inexpensive method for the evaluation of the solution of the lowest order Raviart-Thomas mixed method. *SIAM J. Numer. Anal.*, 22(3):493–496, 1985. pages 189

[34] C. B. Morrey, Jr. *Multiple integrals in the calculus of variations*. Die Grundlehren der mathematischen Wissenschaften, Band 130. Springer-Verlag New York, Inc., New York, NY, 1966. pages 56, 111

[35] I. Muga and K. G. van der Zee. Discretization of linear problems in Banach spaces: residual minimization, nonlinear Petrov–Galerkin, and monotone mixed methods. arXiv:1511.04400v3 [Math.NA], 2018. pages 264

[36] R. H. Nochetto and J.-H. Pyo. The gauge-Uzawa finite element method. I. The Navier-Stokes equations. *SIAM J. Numer. Anal.*, 43(3):1043–1068, 2005. pages 394

[37] J. W. Pearson and A. J. Wathen. A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numer. Linear Algebra Appl.*, 19(5):816–829, 2012. pages 378

[38] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov. Least-Squares mixed finite elements for second-order elliptic problems. *SIAM J. Numer. Anal.*, 31(5):1368–1377, 1994. pages 305

[39] M. J. D. Powell and M. A. Sabin. Piecewise quadratic approximations on triangles. *ACM Trans. Math. Software*, 3(4):316–325, 1977. pages 21

[40] R. Rannacher and S. Turek. Simple nonconforming quadrilateral Stokes element. *Numer. Methods Partial Differential Equations*, 8(2):97–111, 1992. pages 190

[41] I. Smears. Robust and efficient preconditioners for the discontinuous Galerkin time-stepping method. *IMA J. Numer. Anal.*, 37(4):1961–1985, 2017. pages 378

[42] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21(5):995–1011, 1984. pages 454

[43] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Germany, second edition, 2006. pages 358

[44] R. Verfürth. On the constants in some inverse inequalities for finite element functions. Technical report, Ruhr-Universität Bochum, 2004. pages 62

[45] M. Vohralík. Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. *Math. Comp.*, 79(272):2001–2032, 2010. pages 277

[46] W. L. Wendland. Strongly elliptic boundary integral equations. In *The state of the art in numerical analysis (Birmingham, UK, 1986)*, volume 9 of *The Institute of Mathematics and its Applications Conference Series. New Series*, pages 511–562. Oxford Univ. Press, New York, NY, 1987. pages 134